FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Vision Based Context Categorization for All-Terrain Robot

David Alexandre Calado Pereira Chaínho

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Engenharia Electrotécnica e de Computadores.

Orientador: Doutor José António Barata de Oliveira

Presidente do Juri: Doutur José Manuel Matos Ribeiro da Fonseca

Vogais: Doutor José António Barata de Oliveira
Doutor Pedro Alexandre da Costa Sousa

Lisboa
Outubro 2010

UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia

Departamento de Engenharia Electrotécnica

Categorização de Ambientes Baseada em Visão para Robôs Todo-o-Terreno

David Alexandre Calado Pereira Chaínho

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Engenharia Electrotécnica e de Computadores.

Orientador: Prof. José António Barata de Oliveira

Lisboa
2010

UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia

Departamento de Engenharia Electrotécnica

Vision Based Context Categorization for All-Terrain Robots

David Alexandre Calado Pereira Chaínho

Dissertation presented at the Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa to obtain the Master degree in Electrical and Computer Engineering.

Supervisor: Prof. José António Barata de Oliveira

Lisboa
2010

Vision Based Context Categorization for All-Terrain Robots

Copyright © 2010 por David Alexandre Calado Pereira Chaínho e Universidade Nova de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios cientificos e de admitir a sua cópia e distribuição com objectivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor

Acknowledgements

Firstly, I would like to show my honest gratitude towards my dissertation supervisor, Prof. José Barata, for the opportunity, motivation and support to make this happen. Also, I want to give my deep thanks to Pedro Santana for all the comments, knowledge, and valuable help throughout the whole dissertation.

I also would like to thank my parents, Maria and António, my brothers, Marília, and Jorge, and also José Garrido and my girlfriend, Susana Gomes. Finally, I thank my friends and colleagues for all the support, specially, Nelson and Magno.

Learn as if you were going to live forever.

Live as if you were going to die tomorrow.

Mahatma Gandhi

Resumo

Esta dissertação apresenta um modelo que permite a robôs autónomos aprenderem de forma incremental, associações entre o contexto global onde este está imerso e os comportamentos que realiza no ambiente. De certa forma, o robô aprende o que um determinado ambiente lhe oferece, em termos comportamentais (ex.: seguir trilhos, evitar obstáculos). O modelo proposto tem como objectivo ajudar o robô a prioritizar a alocação de recursos percepcionais, e consequentemente, contribuir para focar a sua atenção visual. Para capturar o contexto global, é usado um mecanismo de *gist* para se obter um descritor global do cenário. A focalização na possibilidade de acções oferecidas pelo ambiente em vez de nos objectos presentes no mesmo, isto é, a associação do contexto com acções ou comportamentos invés de com o objecto que activou os mesmos, permite uma aprendizagem auto supervisionada sem a necessidade de se assumir representações simbólicas de objectos, facilitando assim a integração do modelo num sistema em desenvolvimento. O foco nos comportamentos também contribui para a compreensão do papel da coordenação sensório-motora na organização de comportamento adaptativo. Resultados positivos foram obtidos com uma experiência em ambiente natural, que consistiu em transportar uma câmera de vídeo à mão como se fosse carregada por um robô real com um determinado conjunto de comportamentos, tais como seguimento de trilhos, contorno de obstáculos e vaguemento.

Abstract

This dissertation presents a model to allow an autonomous robot to incrementally learn associations between the global context in which it is immersed and the most important behaviours used by the robot in that specific context. In a way, the robot learns what opportunities can a given environment provide in terms of behaviour (e.g., obstacle avoidance, trail following). The proposed model aims at helping the robot prioritising its perceptual resources, and consequently contributes to improve its visual capabilities or skills. In order to capture the global context, a *gist* mechanism is used to obtain a global descriptor of the scene. The focus on affordances, rather than on objects, i.e., associating context with behaviour instead on the objects that activate the behaviours, enables a self-supervised learning mechanism without assuming the existence of symbolic object representations, thus facilitating the integration of the model on a developmental framework. The focus on affordances also contributes to our understanding on the role of sensorimotor coordination in the organisation of adaptive behaviour. Positive results are obtained with a physical experiment in a natural environment, where a handheld camera was transported as if it was being carried by an actual robot with a set of predefined behaviours, such as obstacle avoidance, trail following, and wandering.

List of Symbols and Notations

Symbol	Description
$\mathbf{b}^{\mathbf{l}}[n]$	Behaviour selection vector
$\mathbf{b}^{\mathbf{v}}[n]$	Behaviour classification vector
$\mathbf{g}[n]$	Gist descriptor
m	Newly created memory element
η	Maximum size of a new memory element
ζ	Minumum size of a new memory element
δ	Maximum Chi-Square distance between two gist vectors
ρ	Memory element fusion threshold
$\mathcal{X}^2$	Chi-Square distance

Contents

Acknowledgements **vii**

Resumo **3**

Abstract **5**

List of Symbols and Notations **7**

Contents **9**

List of Figures **11**

List of Tables **13**

1 Introduction **15**

1.1 Problem Statement 17

1.2 Solution Prospect 18

1.3 Dissertation Outline 19

1.4 Further Readings 20

2 State of the Art **21**

2.1 Biological Inspiration 22

2.1.1 Adaptive Behaviour 22

2.1.2 Gist 23

2.1.3 Affordances 24
2.2 Scene classification 25
2.3 Scene Classification and Perception Modulation 28

3 Model **31**
3.1 Gist Calculation 33
3.2 Incremental Learning 33
3.3 Gist Classification 39
3.4 Gist Classification Confidence Level 40

4 Experimental Results **43**
4.1 One Shot Learning Capabilities 46
4.2 Generalisation Capabilities 48
4.3 Prediction Capabilities 51
4.4 Confidence Level - β 53

5 Conclusions, Contributions and Future Work **59**
5.1 Summary of Contributions 59
5.2 Conclusions 60
5.3 Future Work 61

Bibliography **63**

A Additional Results **69**

List of Figures

2.1 Illustration of the the experiment realized in [McVea and Pearson, 2007]. . . . 23

2.2 Dataset images from [Torralba et al., 2003]. 27

2.3 Visual Features used in Gist Model proposed in [Siagian and Itti, 2005]. 28

3.1 Model's Building Blocks. 32

3.2 Example frames with corresponding HSV histogram illustration. 34

3.3 Fluxogram of element creation procedure. 36

3.4 Fluxogram of the insertion of a newly created element into the memory. 37

3.5 Example of memory operation during element merging process. 38

4.1 Experimental environment. 43

4.2 Typical frames labeled as *to be followable*. 45

4.3 Typical frames labeled as *to be avoidable*. 45

4.4 Typical frames labeled as *wanderable*. 45

4.5 Analysis of frame 524. 47

4.6 Analysis of frame 718. 47

4.7 Analysis of frame 939. 48

4.8 Analysis of frame 1582. 49

4.9 Analysis of frame 3157. 50

4.10 Plot of number of elements. 51

4.11 Analysis of frame 5472. 52

4.12 Analysis of frame 5562. 53

4.13 Analysis of frame 5632. 54
4.14 Analysis of frame 5782. 55
4.15 Plots of experimental results. 56
4.16 Elements contained in memory by the end of the run. 57

A.1 Frames from location a in Fig.4.1. 71
A.2 Frames from location a in Fig.4.1. 72
A.3 Frames from location b in Fig.4.1. 73
A.4 Frames from location b in Fig.4.1. 74
A.5 Frames from location d in Fig.4.1. 75
A.6 Frames from location d in Fig.4.1. 76
A.7 Frames from location e in Fig.4.1. 77
A.8 Frames from location e in Fig.4.1. 78
A.9 Frames from location i in Fig.4.1. 79
A.10 Frames from location i in Fig.4.1. 80
A.11 Frames from location j in Fig.4.1. 81
A.12 Frames from location j in Fig.4.1. 82
A.13 Frames from location k in Fig.4.1. 83
A.14 Frames from location k in Fig.4.1. 84
A.15 Frames from location l in Fig.4.1. 85
A.16 Frames from location l in Fig.4.1. 86

List of Tables

4.1 Table with parameter values used in the experiment. 44
4.2 Table with the average times. 46

Chapter 1

Introduction

Environmental context is known to modulate several aspects of animal behaviour, such as its locomotion [McVea and Pearson, 2007]. The importance of context to the animal's survival is so strong that, in the case of humans, there are situations where it is not even possible to consciously suppress its effects altogether [Reynolds and Bronstein, 2004]. Robustness and parsimony in visual search is also known to be strongly correlated with contextual cues [Oliva and Torralba, 2007], in line with active vision research [Ballard, 1991]. Computational models in this case focus on the learning of the statistics describing objects and typical scenes co-occurrence [Torralba et al., 2003]. Thus, the acquisition of this knowledge, according to these models, is based on the existence of a mechanism able of determine if a given object is present in the scene, which is ultimately used to supervise the learning process. However, a global isomorphic representation of the object [Marr, 1982], is unlikely to exist in an embodied agent whose autonomous development occurs bottom-up, in interaction with the environment. Conversely, representations are distributed and purpose-oriented [Goodale, 2008], making a signal to supervise the learning process hard to define.

Under the embodied cognition framework [Pfeifer and Scheier, 1999], perception can only be understood in terms of behaviour, meaning that body, nervous system and environment must be understood in an holistic way [Ashby, 1952, Beer, 1995, Thelen and Smith, 1996]. So, sensorimotor coordination plays a key role on adaptive behaviour [Ballard et al., 1997,

Pfeifer and Scheier, 1999, Mossio and Taraborelli, 2008], and in particular in the shaping of sensory information so as to facilitate perception [Sporns and Lungarella, 2006]. In fact, representations may very well be defined themselves in terms of sensorimotor dynamical states [Scheier et al., 1998, Beer, 2003, Floreano et al., 2004, Nolfi, 2005]. This further complicates the definition of a well localised and steady-state signal to supervise the learning process.

Focused on autonomous mobile robots, and not on general purpose vision systems, this dissertation solves the learning supervision problem by using context to predict affordances [Gibson, 1979, Chemero, 2003], rather than objects. Affordances are possible interactions that the environment offers to an agent equipped with a set of behaviours that allow it to be able to exploit present objects. In this way, it is suppressed the need of explicit object representations. A by-product of this property is the ability to operate even when the control behaviours are yet not fully matured. Since a behaviour is intrinsic to the robot, in the limit, it can be developed in interaction with the environment.

The model starts by assuming that the robot is already capable of exploiting and selecting the most adequate environment's affordance at each decision time. In the presence of a given object, i.e. the aggregate of a given set of perceptual features, the robot knows which behaviour from its predefined set is better applied to it. An example is the *follow* behaviour, which can be effectively applied in the presence of a *trail*. Hence, in the case of a trail, the affordance is *to be followable*. This object-centered knowledge can be evolved [Slocum et al., 2000] or learnt [Fritz et al., 2006, Kim et al., 2006] by having the robot testing its behavioural repertoire in encountered objects. The model is thus operating on a more advanced developmental stage, exploiting the knowledge obtained so far.

The model's second assumption is that the learnt affordances are used to trigger the corresponding behaviours according to a layered behavioural hierarchy [Arkin, 1998]. Hence, the winning behaviour at each moment is associated to the current visual context and stored in the robot's short-term associative memory. Latter on, this memory can be consulted to predict which behaviours are the most appropriate given the visual context at the recalling moment, and by consequence, which affordances are more likely to be present in the environment. Given

the likelihood of a given affordance to occur, the robot should be capable of parsimoniously, and in a context-dependent way, determining how much relevant is to search for a given object, and consequently, how much perceptual resources must be allocated to it. In addiction, with this information, the system may promote the activation of the most relevant behaviours. This enables prediction and stability in face of local environment variations, although this has not been validated in this dissertation.

In the proposed model, visual context is captured through the gist of the scene. The gist is obtained by calculating global statistics of low level features extracted from the visual input. Being a global descriptor, gist is highly fast and robust to local environment variations [Oliva and Torralba, 2007]. This is particularly interesting as it enables the robot to exploit contextual cues robustly and parsimoniously. In addition to reduce sensitivity to varying robot's posture changes, where the scene is observed from different perspectives, the gist provides highly generalisable contextual cues, so enabling their reuse in new environments and robustness facing local environment variations.

The use of gist in the autonomous robotics domain has mostly been limited to learning of places [Siagian and Itti, 2007] and scene categories [Collier and Ramirez-Serrano, 2009] for localisation and mapping purposes. In these works, learning is done off-line and supervised by an external signal (e.g., a symbolic label of the scene). Conversely, our model operates fully online and learning is self-supervised, i.e., the teaching signal is directly available from the behavioural repertoire of the robot.

1.1 Problem Statement

This dissertation contributes to the problem of how the robot can use the knowledge of the environment in which is immersed, to help it in various ways. As mentioned before, this can be useful to boost the selection of a behaviour or to focus robot's perceptual resources to search for a given object. To be able to solve this problem a set of requirements must be met by the model:

1. It must assume that the robot is already equipped with a predefined hierarchy of behaviours. Given a certain task, this hierarchy should be capable of determining what the environment affords and choose the adequate behaviour. This is fundamental because the chosen behaviour will be mapped with the visual context, and this mapping will be learnt by the robot.

2. The proposed model must be robust when encountering local variations in the environment, filter image noise and have good generalisation capabilities for describing the scene. This is important to enable reliable scene classification and therefore, to boost behaviour selection or perceptual resources focusing.

3. The proposed model must be able to run in real time. Otherwise, the model's output could not be used to properly modulate the robot's decisions and allocate perceptual resources. Introducing a big delay in this process would inhibit the potential benefits that the proposed model would bring.

1.2 Solution Prospect

This dissertation proposes the following solutions for the identified problems:

1. To be able to generalise the existing environment, the proposed model considers the gist of the scene as the mechanism to compute environmental context. The gist consists in extracting global information of the robot's visual field to create a signature representing the environment where the robot is [Torralba et al., 2003]. In this study, the gist of the scene is represented by a simple and fast to compute histogram over the whole agent's visual input. Although more accurate methods exist, the simplicity of the model ensures frame rate performance.

2. Lazy learning is utilised to store the association between the global context and the active behaviour. This method enables training data to be stored without further processing

until it is required for classification purposes. This allows to locally approximate the learnt function according to the stored training examples. This enables one-shot learning and therefore fast adaptation. Although being fast, this method requires large memory to store all training examples.

3. Since there is a correlation in time in the robot's input sensory flow, sensor data is highly redundant. The model exploits this properly by grouping similar frames into segments (calculated by their average), which results in memory optimization and noise sensitivity reduction. To reduce even more the impact of noise, small memory elements are discarded from memory because they reflect momentaneous variations in the sensory flow, therefore, noise.

4. The model also incorporates a mechanism to estimate the confidence level in the classification. Among others, this mechanism depends on the matureness of memory elements, where the number of frames that supports these elements is taken into consideration. This way, the degree of modulation of the behavioural hierarchy of the robot can be weighted by the confidence level of the classification.

1.3 Dissertation Outline

This dissertation is organised as follows:

Chapter 2 gives a brief overview of the state of the art regarding biological inspiration, scene categorisation techniques and its applications in autonomous robots;

Chapter 3 describes the full model applied to autonomous mobile robots;

Chapter 4 presents a set of experimental results, where the strengths and weaknesses of the model are analysed as well as a results discussion, in a qualitative point of view;

Chapter 5 gives some conclusions about the developed work and future work possibilities.

1.4 Further Readings

The model proposed in this dissertation has already been published:

[Santana et al., 2010] Santana, P., Santos, C., Chaínho, D., Correia, L. and Barata, J. (2010). Predicting Affordances from Gist. In *Proceedings of the International Conference on Simulation of Adaptive Behaviour (SAB 2010)*, pages 325-334, Paris. Springer.

Chapter 2

State of the Art

This chapter outlines the biological inspiration for this dissertation and the state of the art in gist calculation algorithms. Global context is known to affect animals and humans behaviour [McVea and Pearson, 2007, Reynolds and Bronstein, 2004] and to be useful in perception guidance (Section 2.2). Global context can help the agent in achieving more robust and parsimonious behaviour. Knowledge of the environment the robot is in can be useful to predispose behaviour engagement bringing benefits like improved reaction times and overall stabilization. This is due to the prediction ability that comes from global context analysis. It can also help to stabilise behaviour selection because of the insensitivity of global context to local changes. For example, a robot moving in a cluttered environment passing by an obstacle free area, if only relying in the reactive system, can speed up its locomotion and become less conservative. In this case, it can be dangerous to the robot because, globally, the environment has a high chance of having obstacles (cluttered), so it requires a careful approach that the robot could neglect. However, this idea can also be applied in perception. Knowing what the environment affords to the robot, given a certain task, can help it to guide its visual attention or adapt its sensory sampling frequency.

This chapter starts by presenting an overview of studies done in animals and humans that inspired this dissertation (see Section 2.1). In short, these studies showed that the association between context and behaviour activation can be learnt. An important component of context is the scene category where the agent is immersed in. In the last few years, scene categorisation

has become more important to computer vision and robotics, to help in several tasks. Early studies focused in simple scene categorisation but latter studies evolved to biologically plausible models, such as gist (see Section 2.2). Recent work shows how gist can be applied to modulate perception in robotics, in order to help choosing between perception modules depending on the environment (see Section 2.3).

However, as it will be shown, no previous work links gist with affordances prediction, and consequently with behaviour activation, in particular with online learning. This is essential to enable adaptation in robotics.

2.1 Biological Inspiration

2.1.1 Adaptive Behaviour

The motivation for this dissertation comes from various work that show that animals and humans take environmental context into great importance to modulate several aspects, such as locomotion [McVea and Pearson, 2007]. In some cases, this importance is so great that is not even possible to consciously suppress its effects altogether [Reynolds and Bronstein, 2004].

In the work of Pearson et al. [McVea and Pearson, 2007], cat locomotion is studied. A common feature of locomotion is the the ability of altering its pattern to adapt to various environments with the goal of maintaining stability and efficiency [Pearson, 2000]. A series of experiments realized on cats were made. They consisted in having the cat to walk in a treadmill with a hind back leg, emulating an obstacle striking this leg in its swing phase. The cat reacted to this event with a *hyperflexion*, i.e., quickly lifting the paw up, and over the obstacle (see Fig. 2.1). The other three legs were not obstructed.

The cat was observed walking in other contexts. Notoriously, the *hyperflexion* was not witnessed in these. The persistent *hyperflexion* endured for thousands of undisturbed steps over following days and was only observed on the treadmill, even on the absence of the obstacle.

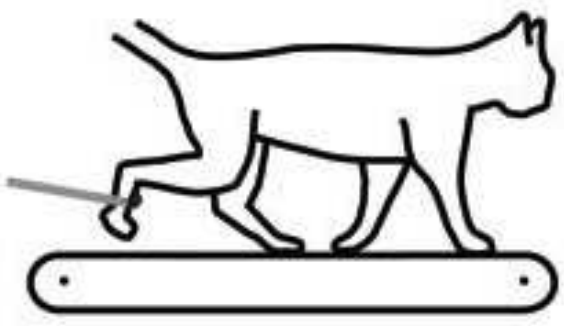


Figure 2.1: Illustration of the stimuli provoked on the cat in the experiment realized in [McVea and Pearson, 2007]. A handheld stick with a padded hook was positioned on the front of the dorsum of the foot throughout sequences of stepping.

This suggests that walking animals may rapidly adapt their locomotion pattern to a specific set of environmental conditions and later apply what it has learnt.

In the work of Reynolds et al. the subjects of study were human beings walking adaptation [Reynolds and Bronstein, 2004]. The experiment consisted on having the subject to walk into a moving platform and adapting to its motion. Once the subject has adapted to walk into the moving platform, one was warned that the platform would no longer be moving in the following trial. Even with this information, the subjects approached the stationary platform at a greater speed than before and a large trunk sway was observed. This after-effect disappeared after three trials, and is representative of action and knowledge dissociation [Reynolds and Bronstein, 2004]. That is, the context suppressed explicit knowledge.

2.1.2 Gist

Through visual perception, humans have the ability of understanding the context of complex original scenes very rapidly, i.e., in about 20 msec [Thorpe et al., 1996], and even if the image is blurred [Schyns and Oliva, 1994]. Early research in scene recognition suggested that the objects encountered by the agent were the visual cues needed to understand the scene. This was refuted by behavioural experiments which concluded that the semantic category of real world scenes can be derived from the analysis of their spatial layout [Oliva and Schyns, 2000]. Furthermore,

studies validate the hypothesis that the processing of the global features and their spatial relationships precede the analysis of local details [Kimchi, 1992]. The global, contextual, and perceptional informations gathered in a glimpse is referred as the *gist* of the scene.

Global features can be summarized as the global statistics of frequency or disposition of local features in a scene. Global features to be considered can be spatial frequency, colour distribution, colour frequency, and disposition of contours.

2.1.3 Affordances

James Gibson was an American psychologist, considered one of the most important 20th century psychologists in the field of visual perception. He was responsible for the creation of the term *affordance*, in his work, The Theory of Affordances [Gibson, 1977]. According to Gibson, an affordance is something rather simple:

> The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. [Gibson, 1977]

Therefore, an affordance ends up being a resource that the environment offers to any animal that has the ability to perceive it and exploit it. Therefore, an affordance varies from animal to animal and environment to environment, and affordances are opportunities provided by the latter if the former is able to exploit them, e.g., trees can be climbed by monkeys but not by cows, so *climbable* is an affordance to the monkey and not to the cow.
The theory of affordances can be useful to robotics and enable to think about the robot's abilities and perception of the environment in a holistic way, and originate new approaches to several problems in robotics such as learning and behaviour.s

2.2 Scene classification

Vision-based scene recognition for mobile robotics is getting more attention from some years ago. Classifying an image to a specific scene category can find application in multiple areas. In image retrieval, it can help organise an image database [Chang et al., 2003]. In autonomous robot navigation, it can help it choose the adequate perception modules and techniques [Collier and Ramirez-Serrano, 2009].

Scene categorisation proves to be a challenge due to the ambiguity and unpredictability of the content of scene images. Complexity increases with illumination, scale and, angle variation. So, in order to efficiently capture the gist of a scene it is necessary to obtain a set of global features. These features are low-level, global descriptors of the image and can vary from colour histograms of various colour spaces to orientation histograms, colour moments (median and standard deviation), curvature histograms, etc. This way we can obtain a global descriptor of an image which will reflect the gist of the environment.

Early approaches on scene classification started by categorising scene images in simple domains such as, indoor vs outdoor. In the work of Vailaya et al. [Vailaya et al., 1999], the classification problem is separated into three smaller ones, indoor vs outdoor, city vs landscape, and forest vs mountain. For each of these problems, an image is represented by a feature vector extracted from the image. Analysed features vary according to the classification problem. In outdoor vs indoor, spatial colour and intensity distribution are used, in city vs landscape distribution of edges is used , and on forest vs mountain global colour distributions and saturation values are used instead. A Bayesian classification method is proposed as the solution for the learning problem. The probabilistic models are estimated during a training phase using a Vector Quantization framework [Gray and Olshen, 1997]. This work shows that for different depths, i.e., from scene categories to places, of classification, different global statistics are utilised. However, the global statistics adequate for each classification problem were obtained through offline analysis. Also the categorisation of images was very simplistic.

A greater number of categories is used in the work of Chang et al. [Chang et al., 2003]. A multimodal type of approach is also used but at a different level. In order to try and emulate human perception as good as possible [Goldstein, 2002], Chang et al. utilises a multi-resolution method. There, images are analysed at different resolutions, coarse, medium, and fine. The main features employed are colour and texture. Depending on the resolution utilised, various statistics of these features can be calculated. Colour statistics include histograms and variance. Textures were categorised in terms of structuredness, orientation and scale (coarseness). Images are classified at two levels. At first, images are classified as one of eight top-level categories, i.e., landscape, people, plants, food, etc. A more precise label is handled in the second level. This ranking method is useful for image retrieval. Although this method of scene classification is accurate, it works offline, so it is not adequate to this dissertation.

A similar work in image retrieval was developed by Torralba et al. [Torralba et al., 2003]. However, the focus there is place recognition to aid object recognition. Knowing the scene of an image can help to narrow down the search for a given object, as well as its location. In this work the image's textural properties and their spatial layout are analysed, as opposed to typical colour histograms. Torralba et al. argues that these work generally well in recognising specific places but do not generalise well to new places.

To compute the texture features, Torralba et al. use a wavelet image decomposition. A steerable pyramid algorithm [Simoncelli and Freeman, 1995] is applied to the intensity (monochrome) image. Contrary to previously presented, this work includes spatial information in the scene descriptor. The mean value of the local features averaged over large spatial regions is used for this purpose. With this operations, a D-dimensional feature vector with large dimensions is obtained. To reduce the size of the vector a Principal Component Analysis (PCA) is applied. Fig. 2.2 shows two images of the used dataset and two others sharing the same global features. These images are generated by modulating noise in order to obtain the same features as the original images.

Figure 2.2: Dataset images and their corresponding noise generated image (from [Torralba et al., 2003]). Noise generated images share the same global features of the original images.

An alternate model is proposed in the work of Siagian et al. [Siagian and Itti, 2005]. The bio-inspired proposed model uses the Visual toolkit by Itti et al. [Itti et al., 2005], featuring a Saliency Model that processes the image through a number of low-level visual ”channels” at multiple spatial scales. The features used by this gist model can be depicted in Fig. 2.3.

In order to reduce the feature vector’s high dimension, a Principal Component Analysis (PCA) is applied followed by an Independent Component Analysis (ICA). The final feature vector’s size is free of irrelevant visual features. The classification is performed by a three-layer neural network, trained with back-propagation algorithm on the reduced gist descriptor. The experimental results are very good, having a very high success rate. The authors suggest that context based vision can aid a mobile robotic in the localisation task.

This work solves the scene categorisation problem with good results. It is a complete solution but it only focuses in classifying images with defined labels. Also, the model is unable to be implemented to run in real-time as it is, because it requires an offline training phase and is computationally intensive. In this dissertation, online and fast gist mechanism and learning

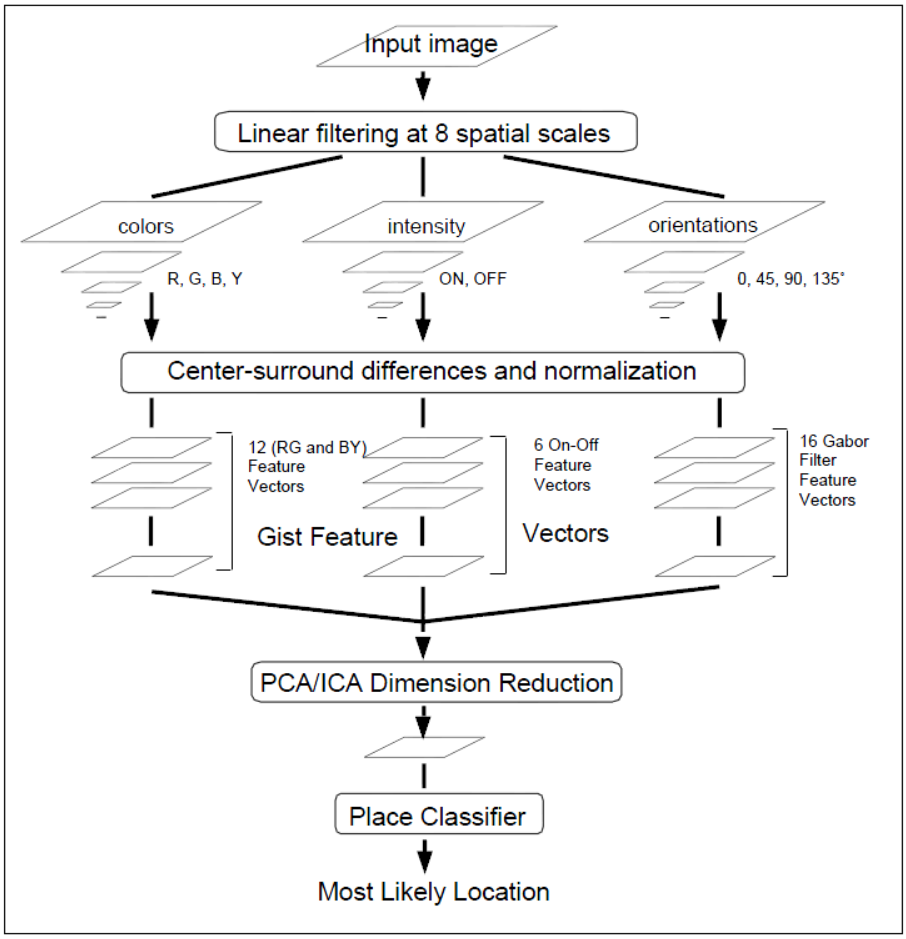


Figure 2.3: Visual Features used in Gist Model proposed in [Siagian and Itti, 2005].

algorithms are required. This requisites are not met by any of the work presented in this section.

2.3 Scene Classification and Perception Modulation

Recent research focused in autonomous robots show the applicability and advantages of environment classification for localisation and mapping, and how this knowledge can aid the robot to allocate perceptual resources according to the surrounding environment.

A work parallel to the one herein present also supports the idea that gist classification can aid robots' decision processes[Collier and Ramirez-Serrano, 2009]. There, the environment is

classified either as outdoor or indoor. This classification helps the robot to choose between perception systems. When operating outdoors, the robot uses GPS and an Inertial Measurement Unit (IMU) to map terrain data. When operating indoors, a 2D Laser based Simultaneous Localisation and Mapping (SLAM) technique is used instead. To learn the mapping between gist and perceptual selection, this work uses Artificial Neural Network (ANN) and Support Vector Machine (SVM), both offline. The training phase generates a classifier to be used online. The best results were got using HSV colour space and ANN learning, where success on switching between perceptual systems according to the type of environment was obtained.

Despite the interesting results, the model presented showed some limitations. The fact of being offline, results in a static system limited to classify according to the labels that were hard coded. Hence, labelling is not done in a selfsupervised manner. Being offline, it also means that the system is not able to autonomously adapt to new environments.

The proposed model in this dissertation aims to provide fast scene categorisation with online learning. These capabilities were not observed in the works presented above. The presented scene classification algorithms, albeit being accurate than the one in the proposed model, do not run in real time. Moreover, these algorithms required heavy off-line learning phases, in opposition to the model presented in this dissertation, which performs rapidly and online.

Chapter 3

Model

Fig. 3.1 illustrates the main building blocks composing the proposed model. As mentioned, the model assumes a bottom layer where a behaviour-based architecture [Arkin, 1998] is responsible for the selection of the affordance to be exploited at each moment, i.e. the behaviour having access to the agent's actuators. The selection among the q possible behaviours is done by a coordinator node, which arbitrates according to a set of fixed priorities. The output of the coordinator node is a binary q-dimensional behaviour selection vector, $\mathbf{b^l}[n]$, whose non-zero element corresponds to the selected behaviour at frame n. This behaviour is selected for actual control of the agent.

On the top of this behavioural architecture, an associative memory grows incrementally so as to learn the mapping between the behaviour selection vector, $\mathbf{b^l}[n]$, and the current visual context, given by the scene's gist, $\mathbf{g}[n]$. The associative memory can be queried at any time for the most likely behavioural selection vector, $\mathbf{b^v}[n]$, given the gist of the current scene, $\mathbf{g}[n]$. Given the global nature of the gist, this prediction is quite often affected by environmental information located in the agent's far field-of-view. This makes the prediction highly useful to modulate the agent's behaviour. One possible exploitation of the associative memory is that, given the current visual context, a behaviour which it is known to be likely to become active could be predisposed. Another possibility, is the allocation of perceptual resources to the detection of this behaviour's associated affordance. With particular utility for the behavioural

modulation aspect, a confidence level on the prediction, $\beta[n]$, is also provided. This enables to access whether predictions are likely to be accurate, and should consequently be considered for the action selection modulation.

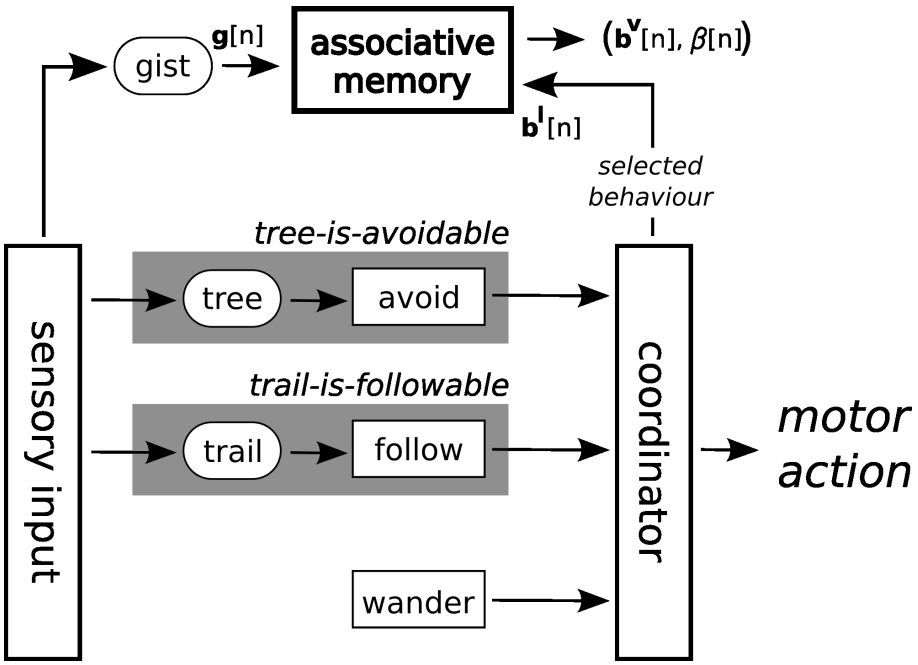


Figure 3.1: Model's Building Blocks. Ovals correspond to object percepts, whose labels are simple descriptions and not symbolic representations. The coordinator basically selects which affordance is exploited at each moment, i.e. whose corresponding behaviour is activated. Gray shadows represent the system's built-in affordances, i.e. the link between a given object and an agent's behaviour.

For the purpose of the current study, a set of two affordances are considered, namely *to be avoidable* and *to be followable*. An example of an object category affording *to be avoidable* is *tree* and of an object category affording *to be followable* is *trail*. A found trail is thus assumed to be followed by the agent. Avoidable objects are more relevant for the agent's survival and so their presence right in front of the agent subsume the other affordance. In the absence of any of these in the environment, the agents starts wandering. When wandering, the presence of any avoidable object, be it on the front of the agent or not, will activate the avoidance behaviour.

3.1 Gist Calculation

In this study, the environmental context is defined in terms of the scene's visual gist.

Although more complex and accurate methods exist, that use a wide range of features, [Torralba et al., 2003, Siagian and Itti, 2007, Collier and Ramirez-Serrano, 2009], in this study the gist of the scene is represented by a simple and fast to compute histogram over the whole agent's visual input. Being a global descriptor, the gist is not sensitive to local variations on the environment. This in turn results in good generalisation capabilities in categorising the scene, which as experimental results will show, help the agent when facing new environments.

Concretely, the gist descriptor, $\mathbf{g}[n]$, is a three dimensional histogram obtained from the whole image in the HSV colour space that provides the best results for scene classification as shown by Collier et al. [Collier and Ramirez-Serrano, 2009], where he compares between several colour spaces, RGB, LUV, and HSV. To reduce sensitivity to illumination effects, the saturation (S) and value (V) channels are represented by only 4 bins, whereas the hue (H) is represented by 16 bins. This descriptor is consequently a vector of 256 numerals whose combined values are representative of a given type of environment, such as *forested*. Illustration of HSV colour space can be found in Fig. 3.2.

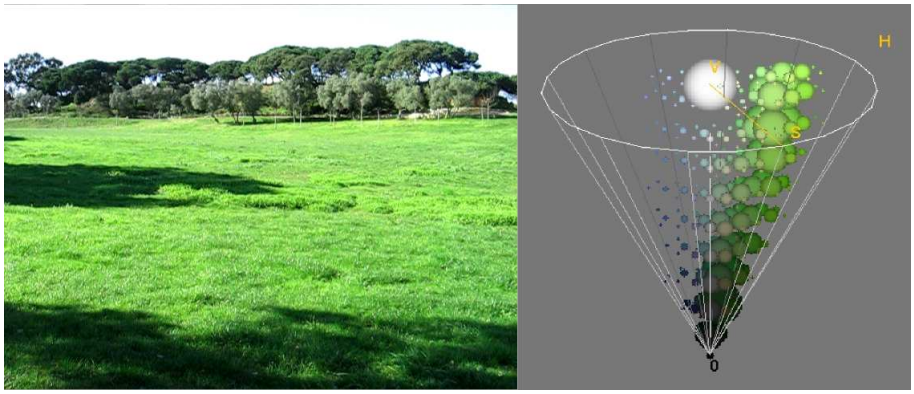
Note that no label is associated to the descriptor. As it will be shown, the learning process just associates this non-symbolic descriptor to behaviour selections taken by the agent.

3.2 Incremental Learning

Once the gist is computed, it can be associated to the selected behaviour being engaged by the agent. This association can then be exploited to know which behaviour should be acted given the current gist, or in other words, which affordance that is more likely to be found in the environment should be attended first.

Most gist-related research has been focused on offline learning, with heavy algorithms, [Torralba et al., 2003, Siagian and Itti, 2007, Collier and Ramirez-Serrano, 2009], which is not

(a) Left: Frame 3160, Right: HSV histogram representation.



(b) Left: Frame 3964, Right: HSV histogram representation.

Figure 3.2: Example frames with corresponding HSV histogram illustration.

adequate for a truly autonomous agent. In this study, the learning procedure follows the lazy learning paradigm, where the training examples are stored until they are necessary, i.e. when recalling is taking place. The biggest advantage of lazy learning is the possibility of locally approximating the learnt function according to the stored training examples. In the limit, a single example is necessary to generate a classification. This enables one-shot learning and consequently fast adaptation. In turn, the biggest disadvantage of lazy learning is the large memory requirements to store all training examples. However, as sensory flow in an embodied agent is highly correlated in time, a large redundancy is observed.

We exploit the existing redundancy on the sensory flow by creating segments of sequential frames, whose first element's gist is similar to the gist of the remaining ones. That is, a segment is created by accumulating frames until the gist descriptor of the current frame is too dissimilar from the one of the first segment's frame, or until an upper bound of η frames is reached. Two

gist vectors are assumed to be dissimilar if the Chi-Square distance between them is above δ. To reduce sensitivity to noise, a newly created segment is rejected from further processing if represented by less than ζ frames. A diagram demonstrating this procedure can be seen at Fig. 3.3.

The average gist, given by

$$\mathbf{s}(m) = \sum_{j=a(m)}^{n} (\mathbf{g}[j]/(n - a(m))) \tag{3.1}$$

of the newly created segment m, is associated to the histogram of behavioural selections occurred during the segment's composing frames,

$$\mathbf{h}(m) = \sum_{j=a(m)}^{n} \mathbf{b}^{\mathbf{l}}[j] \tag{3.2}$$

where $a(m)$ is the index of the segment's first frame and n the index of the current and consequently segment's last frame. The tuple $\langle \mathbf{s}(m), \mathbf{h}(m) \rangle$ is introduced to the associative memory M as shown by Fig. 3.4.

If the average gist of the new segment, $\mathbf{s}(m)$, is significantly similar to the most similar segment already present in the associative memory,

$$\mathcal{X}^2(\mathbf{s}(m), \mathbf{s}(o)) < \rho \tag{3.3}$$

with

$$o = \arg\min_{b \in M}(\mathcal{X}^2\,(\mathbf{s}(m), \mathbf{s}(b))) \tag{3.4}$$

then both are blended, where $\mathcal{X}^2(\cdot)$ is the Chi-Square distance. Otherwise m is simply appended to the memory. Merging occurs by averaging both gist descriptors,

$$\left(\frac{n_m}{n_m + n_o} \mathbf{s}(m) + \frac{n_o}{n_m + n_o} \mathbf{s}(o) \right) \tag{3.5}$$

weighted by their number of supporting frames, n_m and n_o. The behaviour selection histograms

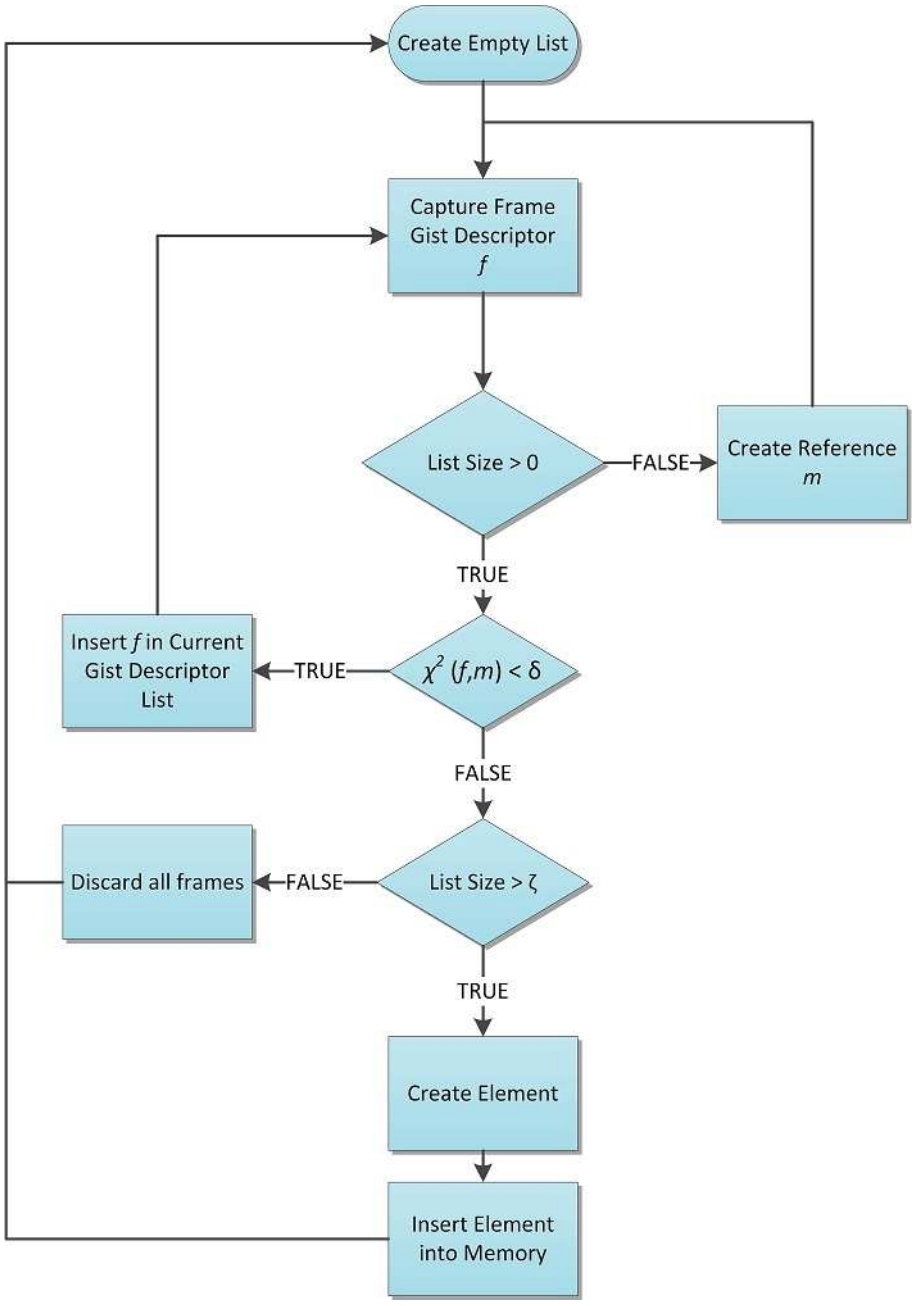


Figure 3.3: Fluxogram of element creation procedure.

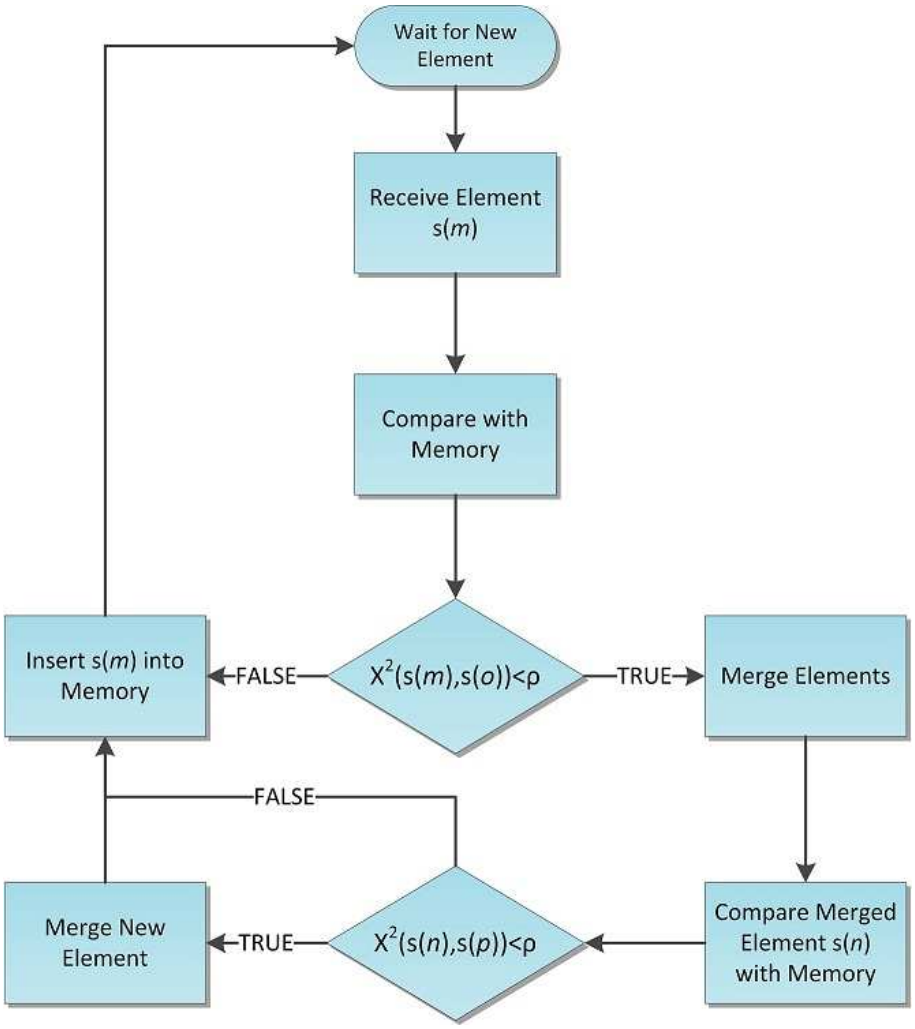
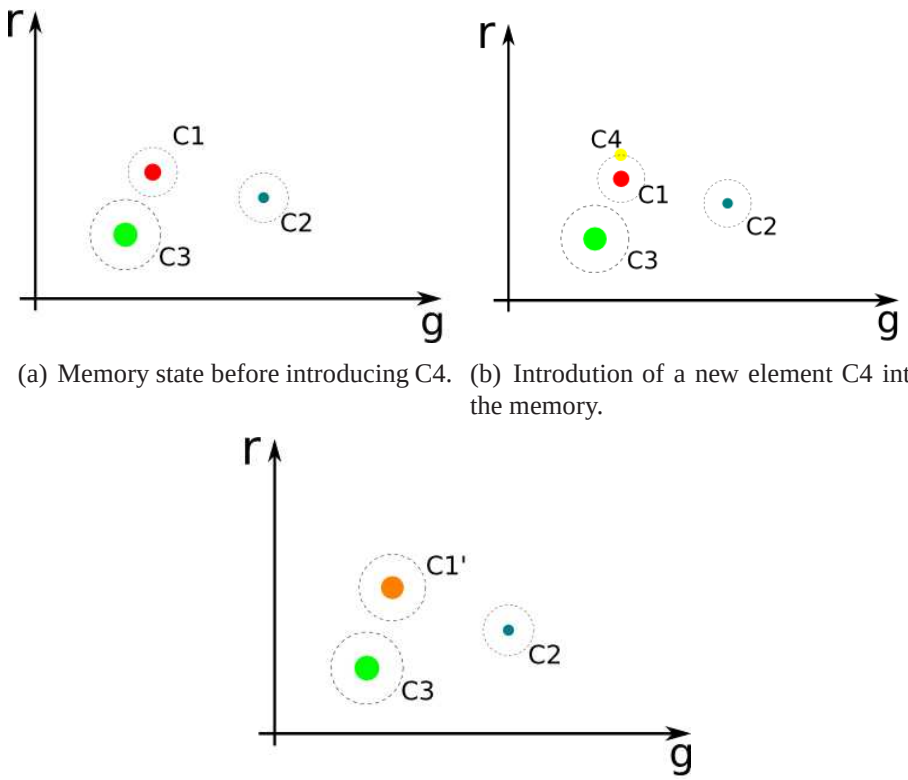


Figure 3.4: Fluxogram of the insertion of a newly created element into the memory.

are also blended via a simple summation, $(\mathbf{h}(m) + \mathbf{h}(o))$. The resulting merged segment is then compared to the second most similar segment to m,

$$p = \arg \min_{b \in M \setminus \{o\}} (\mathcal{X}^2(\mathbf{s}(m), \mathbf{s}(b))) \tag{3.6}$$

and if the merging conditions are met (see above), both segments are merged. This two-step merging procedure is an attempt to avoid the associative memory from growing unbounded, without incurring in excessive processing.



(a) Memory state before introducing C4.

(b) Introdution of a new element C4 into the memory.

(c) New memory state with C1' resulting from C1 and C4 fusion.

Figure 3.5: Example of memory operation during element merging process. Filled circles C1, C2, and C3 are existing elements in the memory, whose gist vector are simplified to be 2-D. The radius of each circle is proportional to the matureness level of the correspoding element. The outer circles that surrounds each element corresponds to the threshold for element merging, that is, when a new element appears within the basin of an outer circle, it is merged with corresponding element. The radius of the outer circle is given by the empirically defined scalar ρ (see Equation 3.6).

An illustration of the clustering process is depicted in Fig. 3.5. In Fig. 3.5(a), the presence of

3 elements in memory can be observed. At Fig. 3.5(b), a newly introduced element in memory, C4, can be seen. That element is similar to C1 and whose Chi-Square distance to it is lesser than ρ, so they can merged. The resulting element of the merging can be seen at Fig. 3.5(c), which new location is the mass centre of both elements, and its new size is the sum of both element's number of supporting frames.

3.3 Gist Classification

Every time a new image frame is obtained, the associative memory can be queried for the most likely behaviours given the gist descriptor, $\mathbf{g}[n]$, of the current scene. This is done according to an adaptation to the weighted k nearest-neighbour method, where $k = 4$ has shown to provide the best results for the tested data-set, i.e. trade-off between accuracy and generalisation capabilities.

In more detail, given the query $\mathbf{g}[n]$, the associative memory is searched for the closest k segments, which are said to compose the ordered set $K = \{m_0, \dots m_k\}$. The order is given by the Chi-Square distance to the query, at the gist descriptor level, i.e.

$$\mathcal{X}^2(m_i, \mathbf{g}[n]) > \mathcal{X}^2(m_j, \mathbf{g}[n]), \forall i > j \tag{3.7}$$

The return to the query, i.e. the classification, is a normalised behaviour selection histogram resulting from the weighted sum of the behaviour selection histograms of the segments in K,

$$\mathbf{b}^{\mathbf{v}}[n] = \sum_{l=0}^{k} \frac{\mathbf{h}(m_l)}{k} w(m_l) \tag{3.8}$$

The weight of a segment $m_l \in K$ is as large as the Chi-Square distance to the query gist descriptor is small, and as high as it is its order in K,

$$w(m_l) = \frac{2 - \mathcal{X}^2(\mathbf{s}(m_l), \mathbf{g}[n])}{2l} \tag{3.9}$$

The magnitude of the elements composing $\mathbf{b^v}[n]$ represent the likelihood of each behaviour to occur, given the current gist, and consequently the possibility of finding their associated affordances.

3.4 Gist Classification Confidence Level

Aside the estimate behaviour selection histogram $\mathbf{b^v}[n]$, the associative memory also returns a confidence level, $\beta[n]$, on the classification. $\beta[n]$ varies according to:

1. The confidence the system has on the visual context, given by ξ;
2. The discrepancy between the predicted and current behavioural context, $d(\mathbf{b^v}[n], \mathbf{b^l}[n])$, which controls the value of γ;
3. The rate of variation of ξ, i.e. $\dot{\xi}$, in case it decreases.

To account for these aspects, β is modelled as,

$$\dot{\beta} = (\xi - \beta) \cdot \alpha_1 + \left(\mathcal{H}(-\dot{\xi})\dot{\xi}\beta\right) \cdot \alpha_2 - (\gamma\beta) \cdot \alpha_3 \tag{3.10}$$

where

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{3.11}$$

and $\mathcal{H}(\cdot)$ is the Heaviside step function. The use of dynamical systems to calculate β assures robustness in the final system.

The higher the number of frames supporting $\mathbf{b^v}[n]$, $f(\mathbf{b^v}[n])$, the more confident is the system on its prediction. This confidence, ξ, is given by

$$\xi = \mathcal{G}\left(f(b^v)[n]\right) \tag{3.12}$$

where

$$\mathcal{G}(x) = e^{-5e^{-\lambda x}} \tag{3.13}$$

is the Gompertz function such that $\xi \in [0, 1]$. This function makes ξ converge faster towards near 1 in face of reliable information from the associative memory, but more slowly in reaching the final value of 1, which would mean that confidence in the predicted behaviour selection vector is maximum.

The more discrepancies exist between the predicted and current behavioural contexts, the less confident the system is on the former, and β should approach zero. This discrepancy, $d(\mathbf{b^v}[n], \mathbf{b^l}[n])$, is defined in terms of the Euclidean distance between both vectors $\mathbf{b^v}[n]$ and $\mathbf{b^l}[n]$. The following dynamical system takes into account these discrepancies,

$$\dot{\gamma} = (1 - \gamma) \cdot d(\mathbf{b^v}[n], \mathbf{b^l}[n]) \cdot \kappa_1 getCluster - \kappa_2 \gamma \tag{3.14}$$

Discrepancies, $d(\mathbf{b^v}, \mathbf{b^l})$, are accumulated in γ at a rate of $k_1 \cdot (1-\gamma)$. γ tends to zero in case there are no discrepancies, meaning one should increase the confidence in the visual context and β should approach γ, that is 1. Similarly, γ tends towards one in case discrepancies are maximum, meaning one should reduce the confidence in the prediction, and β should approach 0.

Chapter 4

Experimental Results

To validate the proposed model, an experiment with the objective of demonstrating the ability of the associative memory to learn a generalisable gist-affordance mapping was carried out. A video composed of 9000 frames, with a resolution of 320x240, (recorded at 15fps) was obtained by a person walking for approximately 10 minutes through a predefined course in a natural park (see Fig. 4.1) with a hand-held camera at the shoulder's height.

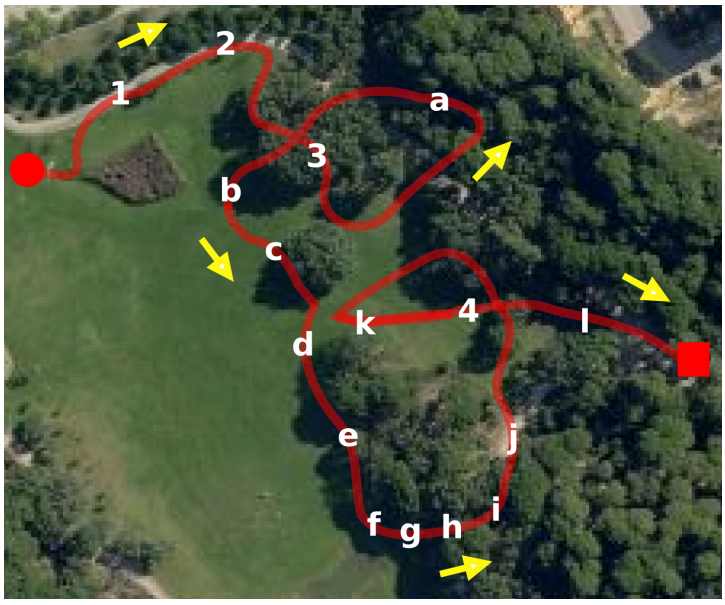


Figure 4.1: Experimental environment. The line corresponds to the motion path, whose direction is cued by the arrows. Letters are key locations, whose associated frames are exhibited in Fig. 4.15.

During image acquisition, the camera felt a considerable level of oscillations, typical in off-road robots. The resulting sudden viewpoint changes and induced blur are stringent conditions with which the model must be able to handle.The camera was moved as similar as possible as it would be if mounted on a mobile robot acting according to the behavioural hierarchy presented in Section 2. That is, when the person selected to follow a trail, the camera was pointed towards its vanishing point (see Fig. 4.2). Any obstacle faced by the person was circumnavigated, thus emulating the avoidance behaviour (see Fig. 4.3). In the absence of a trail and facing obstacles, the person engaged on a wandering behaviour (see Fig. 4.4).

The video was then hand-labeled with respect to which behaviour was being emulated by the person at each frame. That is, the signal that would be output by the behavioural hierarchy, $\mathbf{b}^{\mathbf{l}}[n]$, was manually defined according to the emulated behaviour. The system was then evaluated as if the video was being obtained on-line and $\mathbf{b}^{\mathbf{l}}[n]$ was being generated by the behavioural hierarchy.

The model was implemented with the support of the library OpenCV for low-level computer vision routines. The model ran in an Intel Core 2 Duo 2.8GHz. The system was parameterised as seen in Table 4.1.

Parameter Value	Description
$\zeta = 20$	Minimal Element Creation Size
$\eta = 50$	Maximum Element Size
$\delta = 0.2$	Element Creation Threshold
$\rho = 0.4$	Element's Fusion Threshold
$\lambda = 0.03$	Gompertz Function Weight
$\alpha_1 = 0.6$	Confidence Level Function Parameter
$\alpha_2 = 0.3$	Confidence Level Function Parameter
$\alpha_3 = 0.1$	Confidence Level Function Parameter
$k_1 = 0.8$	Discrepancy Level Function Parameter
$k_2 = 0.2$	Discrepancy Level Function Parameter

Table 4.1: Table with parameter values used in the experiment.

Table 4.2 shows the computation time of each component of the model.

The total average time is 21.67 ms, which results in a processing rate of 46Hz. Being fast is key for this algorithm, as it intends to modulate the behavioural hierarchy. A slow processing

Figure 4.2: Typical frames labeled as *to be followable*.

Figure 4.3: Typical frames labeled as *to be avoidable*.

Figure 4.4: Typical frames labeled as *wanderable*.

Algorithm	Average Time (ms)	Standard Deviation (ms)
Image Processing	19.75	1.37
Database Update	0.76	4.12
Classification	1.16	0.25
Total	21.67	4.34

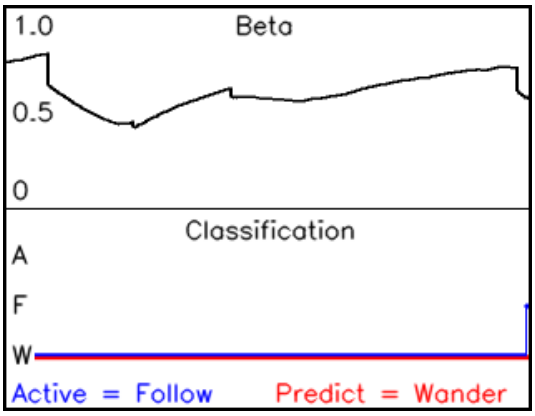
Table 4.2: Table with the average times and their standard deviation, of the various functions of the model.

rate would increase the reaction time of the robot, rendering unreasonable the purpose of fast scene categorization to aid in the robot's decision.

The following sections present an analysis of the field trial. A set of key situations are selected for analysis in order to show the model's qualities and weaknesses. Then, an explanation and discussion of the results is provided.

4.1 One Shot Learning Capabilities

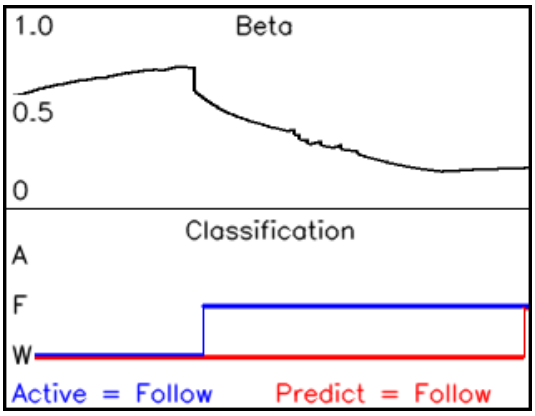
The one-shot learning capability of the system can be appreciated at location `1` (see Fig. 4.1), i.e., soon after the onset of the first trail following. For about 500 frames since the onset of the run, the system gains experience of the environment which affords *wanderable*. At frame 524 the system enters a in a new gist which affords *followable* (see Fig. 4.5). However, the classification remains at *wanderable* because the system is immature and the only elements it contains in memory are elements associated with *wanderable*. At frame 718, i.e, 13 seconds after, (see Fig. 4.6), pinpointed with location `1` in Fig. 4.1, the associative memory was already able to recognise the scene as containing elements *to be followable* (see in Fig. 4.15).



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 524.

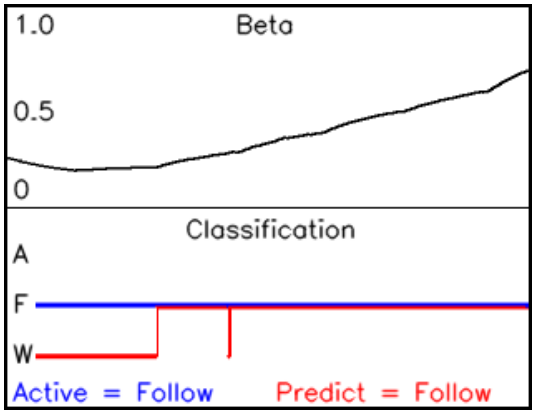
Figure 4.5: Analysis of frame 524.



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 718.

Figure 4.6: Analysis of frame 718.



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 939.

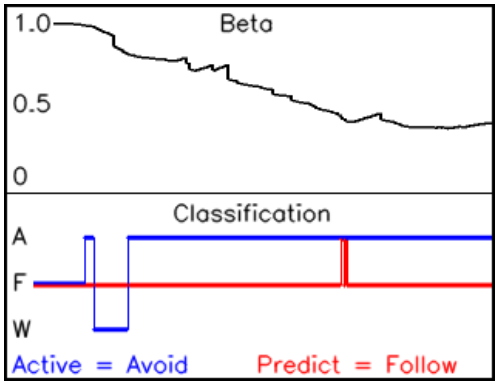
Figure 4.7: Analysis of frame 939.

4.2 Generalisation Capabilities

An example of generalisation is the one depicted in Fig. 4.9, where the associative memory confirms the behavioural hierarchy in what regards the presence of the *to be followable* affordance, and further generalises it by also predicting the occurrence of the *to be avoidable* affordance. This generalisation is boosted by the similarity of the environment in question with the previously experienced one at location 3 (see Fig. 4.8 and Fig. 4.1), where the dense presence of trees induced the behavioural hierarchy to select the *to be avoidable* as the affordance to be exploited (see Fig. 4.15).

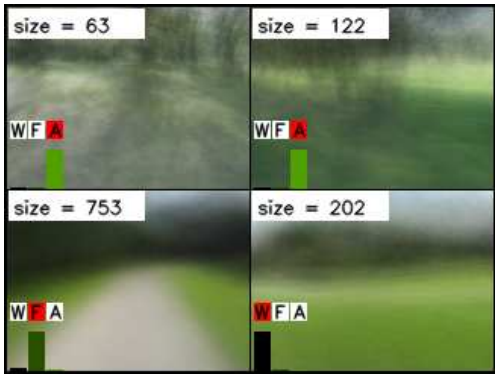
This is a situation where the robot is on a trail and could potentially move at a faster pace. However, there is a nearby presence of obstacles (trees), reflected in the similar evidence in classification output of *to be avoidable* and *to be followable*. Therefore, this could inhibit the robot from increasing locomotion speed and make it adopt a defensive behaviour.

This generalisation ability can also produce erroneous results, as it is the case from frame 1300 to frame 1600 (see Fig. 4.8). Nevertheless, as we can see in the elements window (see Fig. 4.8(a)), the nearest neighbour is correctly identified. However with four nearest-neighbour, $k = 4$, the next most similar memory elements, that are wrong, end up having a considerable

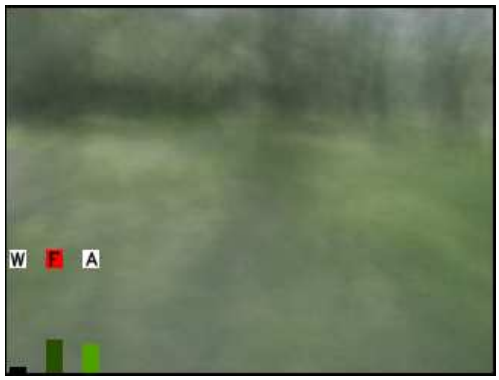


(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 1582.

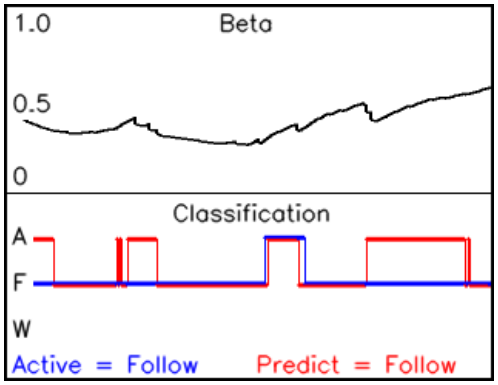


(c) Elements window with the 4-nn elements and their behaviour selection histograms. Upper left image is $k = 1$, lower right is $k = 4$. Behaviour histogram: left bar is *wander*, middle bar is *follow*, right bar is *avoid*.



(d) Weighted average image of all frames used to compute the element in question. Weights defined with equation 3.9. These images are for presentation purposes only; they are not maintained in the system.

Figure 4.8: Analysis of frame 1582.



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b^l}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b^v}[\mathbf{n}])$.

(b) Image correspondent to frame 3157.

(c) Elements window with the 4-nn elements and their behaviour selection histograms. Upper left image is $k = 1$, lower right is $k = 4$. Behaviour histogram: left bar is *wander*, middle bar is *follow*, right bar is *avoid*.

(d) Weighted average image of all frames used to compute the element in question. Weights defined with equation 3.9. These images are for presentation purposes only; they are not maintained in the system.

Figure 4.9: Analysis of frame 3157.

weight in the final result. This is mostly because these elements are very mature, i.e., they have a high number if frames. Moreover, their gist vectors are not dissimilar enough to have a low impact in the classification,e.g., green is strongly present in almost all frames.

The stabilisation of the associative memory happens roughly at half of the run, with a total of 24 elements. This small quantity of elements shows that the model generates a bounded/parsimonious representation of the environment. This is a demonstration of the model avoiding to over-fit the environment, which in turn is one of the causes for its good generalisation ability.

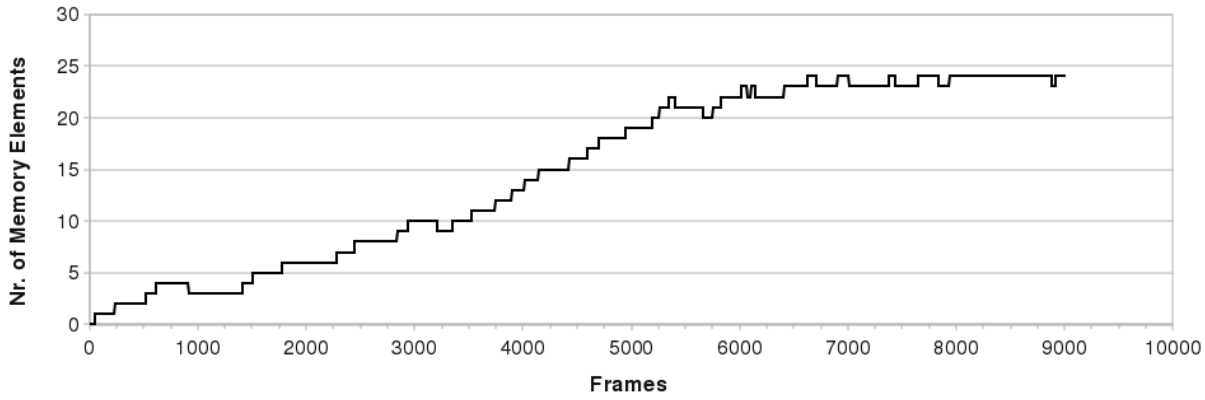
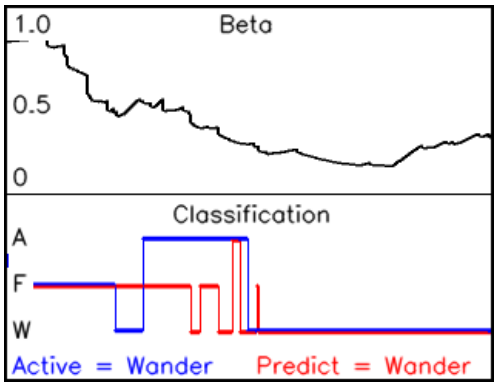


Figure 4.10: Plot of number of elements.

4.3 Prediction Capabilities

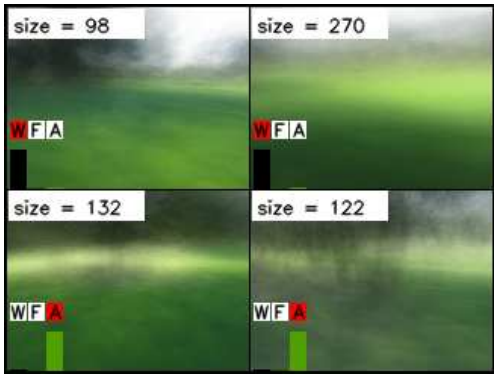
The frame sequence from Fig. 4.11 to Fig. 4.14 illustrates a situation where the behavioural hierarchy switches from wander behaviour to trail following. At the beginning of this sequence (see Fig. 4.11), we can see the behavioural hierarchy decision being *to be wanderable*, and the prediction confirming it. As a consequence, confidence also increases. In frame 5562 (see Fig. 4.12), we can see that the system began to predict *to be avoidable*, which was confirmed by the ground truth some frames after. We can also see in the classification results (see Fig. 4.12(d)) that the classification is divided between *to be wanderable* and *to be avoidable*, which is reasonable as there is some open field ahead with some side trees. Again, with this information, the behavioural hierarchy that could cause the robot's behaviour to be more conservative. At frame 5632 (see Fig. 4.13) we can see that the system starts predicting *to be followable* with a

good confidence factor though with a mix of *to be avoidable*. We suspect that the system learnt that the presence of trees and shadows induce the presence of a trail and also obstacles, which ends up to be true because in the tested environment trails and obstacles co-occur most of the time, i.e. trails have trees nearby.



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

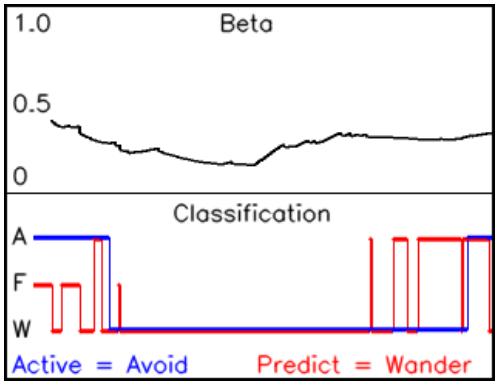
(b) Image correspondent to frame 5472.



(c) Elements window with the 4-nn elements and their behaviour selection histograms. Upper left image is $k = 1$, lower right is $k = 4$. Behaviour histogram: left bar is *wander*, middle bar is *follow*, right bar is *avoid*.

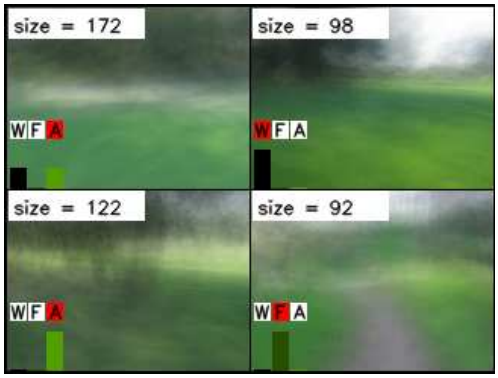
(d) Weighted average image of all frames used to compute the element in question. Weights defined with equation 3.9. These images are for presentation purposes only; they are not maintained in the system.

Figure 4.11: Analysis of frame 5472.



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 5562.



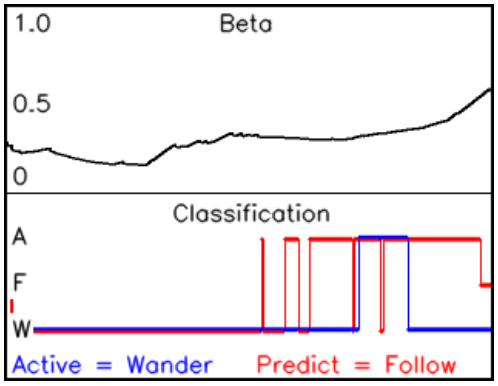
(c) Elements window with the 4-nn elements and their behaviour selection histograms. Upper left image is $k = 1$, lower right is $k = 4$. Behaviour histogram: left bar is *wander*, middle bar is *follow*, right bar is *avoid*.

(d) Weighted average image of all frames used to compute the element in question. Weights defined with equation 3.9. These images are for presentation purposes only; they are not maintained in the system.

Figure 4.12: Analysis of frame 5562.

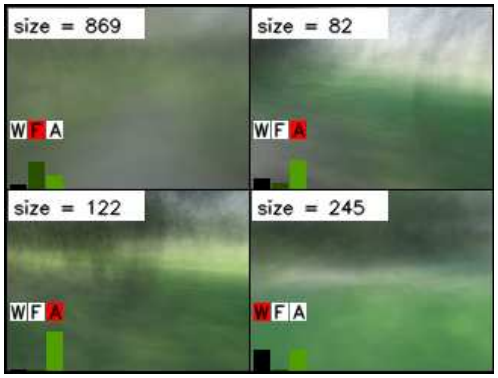
4.4 Confidence Level - β

The plot in Fig. 4.15 shows that when the prediction is stable and it matches the current behaviour selection vector, such as at location 2 (see Fig. 4.1), β is high. Conversely, when the prediction changes often in a short time and consequently mismatches the current behaviour selection vector, as at location 4, β decreases considerably. As a consequence, β shows to be a good indicator of how much certain are the predictions generated by the associative memory. It also can be depicted in Fig. 4.6 that the confidence factor β begins to decrease once classifi-



(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 5632.

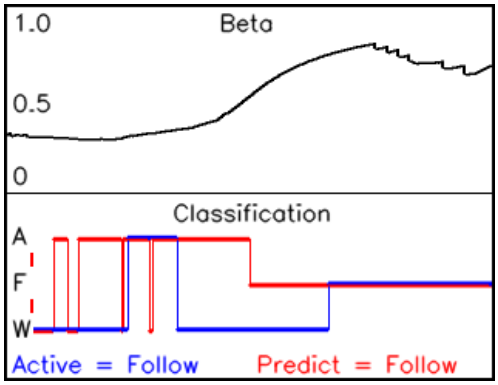


(c) Elements window with the 4-nn elements and their behaviour selection histograms. Upper left image is $k = 1$, lower right is $k = 4$. Behaviour histogram: left bar is *wander*, middle bar is *follow*, right bar is *avoid*.

(d) Weighted average image of all frames used to compute the element in question. Weights defined with equation 3.9. These images are for presentation purposes only; they are not maintained in the system.

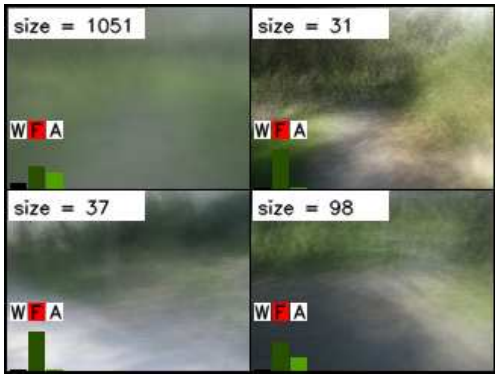
Figure 4.13: Analysis of frame 5632.

cation starts failing and it begins to acquire new elements. These have a low number of frames and, therefore, classification begins to diverge. Once it begins to correctly classify the current gist and elements supporting the classification increase in size, β begins to increase, reflecting a more trustful classification (see Fig. 4.7).

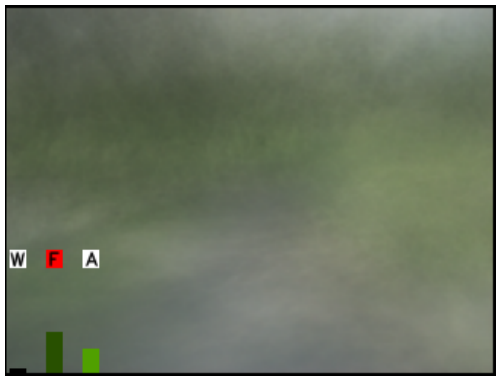


(a) Temporal series, where only the last 320 frames are represented in the plots. Upper Half: β Value. Lower Half: Classification history, blue line is the active behaviour, i.e, $\max(\mathbf{b}^{\mathbf{l}}[\mathbf{n}])$, and red line is the predicted behaviour, i.e., $\max(\mathbf{b}^{\mathbf{v}}[\mathbf{n}])$.

(b) Image correspondent to frame 5782.



(c) Elements window with the 4-nn elements and their behaviour selection histograms. Upper left image is $k = 1$, lower right is $k = 4$. Behaviour histogram: left bar is *wander*, middle bar is *follow*, right bar is *avoid*.



(d) Weighted average image of all frames used to compute the element in question. Weights defined with equation 3.9. These images are for presentation purposes only; they are not maintained in the system.

Figure 4.14: Analysis of frame 5782.

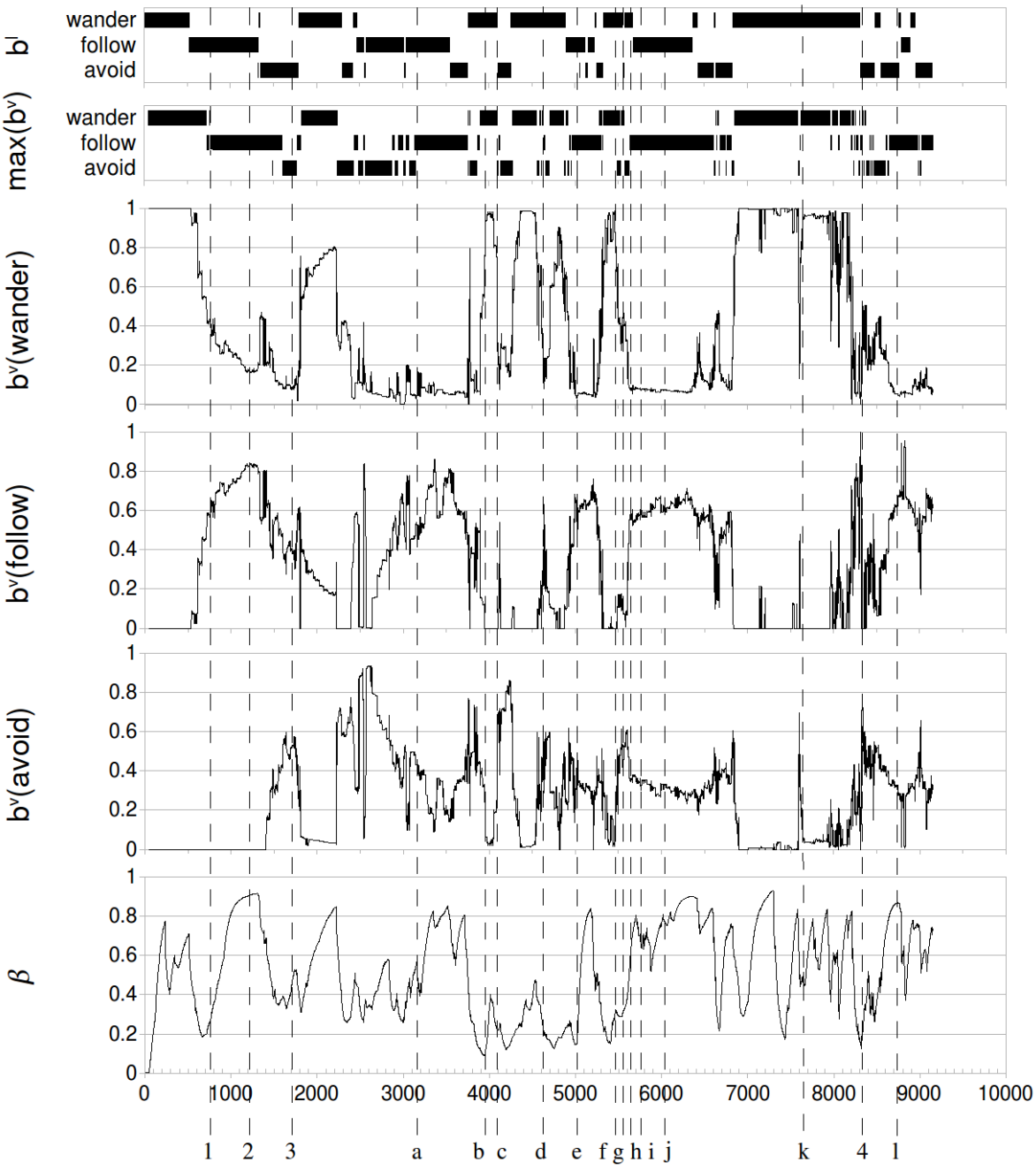
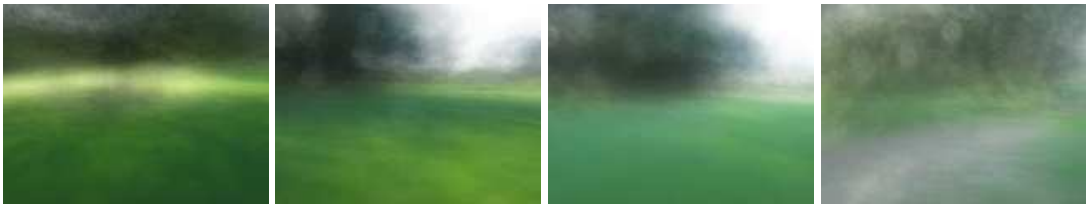


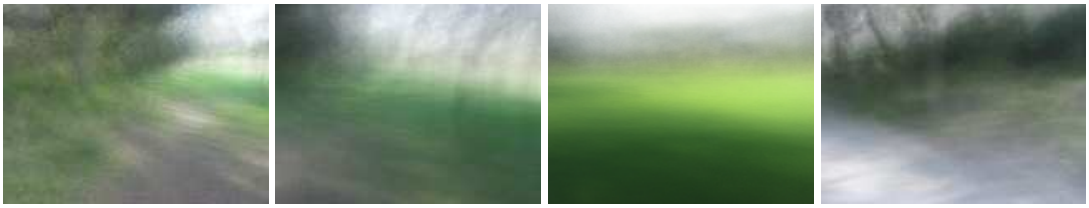
Figure 4.15: Plots of experimental results. The black bars on the $\mathbf{b}^{\mathbf{l}}$ plot correspond to selection made by the behavioural hierarchy at each frame. The bars on the plot underneath, i.e. $\max(\mathbf{b}^{\mathbf{v}})$, refer to the behaviour whose activity is predicted as the strongest one. The three plots bellow are the predicted activity of each behaviour. Labels 1-4 and a-l indicate key locations.

(a) Element 0 (b) Element 1 (c) Element 2 (d) Element 3

(e) Element 4 (f) Element 5 (g) Element 6 (h) Element 7



(i) Element 8 (j) Element 9 (k) Element 10 (l) Element 11



(m) Element 12 (n) Element 13 (o) Element 14 (p) Element 15

(q) Element 16 (r) Element 17 (s) Element 18 (t) Element 19

(u) Element 20 (v) Element 21 (w) Element 22

Figure 4.16: Elements contained in memory by the end of the run.

Chapter 5

Conclusions, Contributions and Future Work

This chapter summarises the dissertation, discusses the proposed approaches and respective contributions, as well as the results obtained, followed by some aspects to be taken into consideration in future work.

5.1 Summary of Contributions

In this dissertation, an incremental learning mechanism used to create associations between the most useful behaviour afforded by the environment and the environment's gist (i.e., visual context) was proposed. The proposed model can help the robot prioritising its perceptual resources on those aspects of the environment that are simultaneously more likely to occur and useful to the robot. It can also be used to predispose behavioural engagement or to stabilise behaviour selection. This was enabled by the ability of gist to provide good generalisation and robustness to local variations. Good generalisation is key to the robot when facing new environments.

The proposed model is self-supervised by nature, enabling its inclusion on a developmental

framework. This is possible because the learning process is focused on what behaviours are afforded by the environment, rather than on its objects and thus does not assume the existence of symbolic object representations. By having behaviour activation information available, labels for learning are obtained straightforward.

A confidence factor β was developed to reflect the solidity of the classification and the weight it should have if used to modulate the behavioural architecture.

The proposed model shows that although context-based visual attention might seem to be a perceptual problem, higher levels of autonomy are more easily obtained if it is seen instead as a sensorimotor problem.

5.2 Conclusions

Experimental results show the ability of the proposed model to properly generalise and predict environments. The lazy learning paradigm used in the model shown to provide one-shot capabilities as a result of its ability to approximate locally the classification function. The model's generalisation benefits can be seen in the experimental results, where it was shown a situation where this capability can be useful to stabilise robot behaviour. This was a situation where the robot might decide to increase its locomotion speed, as it was on a trail. The trail was surrounded by trees rendering this behaviour dangerous. But since the model also reported the avoidance behaviour, the behavioural hierarchy could take this into account and adopt a reduced locomotion speed. The final number of memory elements in the trial field stabilised at 24 elements at nearly half of the run. This shows that the model does not try to overfit the environment, what demonstrates its generalisation properties.

The model's prediction ability was also shown in some situations where the it was anticipated the occurrence of certain affordances that later proved to be present. This could be used by the robot to predispose the corresponding behaviour in a predictive and stable way.

The confidence factor β demonstrated to be an adequate indicator of classification reliability. Finally, the computation times were shown to meet the frame rate requirement. The obtained frame rate of 46Hz adds minimal computation overhead to an existing system.

5.3 Future Work

Despite the promising results, further experiments are required to fully assess the impacts of the model. Concretely, the model must be tested on a real robot. The actual benefits of modulating visual attention with the output of the proposed model must be thoroughly assessed. The process of doing it must also be analysed. For instance, it is necessary to understand how strongly the modulatory signal must be taken into account by the behavioural hierarchy. The way the model should influence the perceptual resources allocation process must also be studied. For example, object detection algorithms activation frequency can be modulated by the classification output, where a more likely object detection algorithm has more resources allocated. It should be noted that the system can only learn at the instant all algorithms provide an output.

Additional future work research lines include:

- Adding temporal filters to the classification output for improved stability.
- Testing the model on a developmental framework, where affordances are being discovered, exploited, and refined, at the same time context is being taken into account.
- Improving the model's gist and learning algorithm. Besides colour histograms, gist computation may also include orientation and curve histograms as well as spatial info [Torralba et al., 2003]. In the learning algorithm, more advanced online learning algorithms [Atkeson et al., 1997] can be studied and introduced into the model.
- Exploiting other sensory modalities such as sound, as well as temporal information for wider context assessment.

Bibliography

[Arkin, 1998] Arkin, R. C. (1998). *Behavior-Based Robotics*. The MIT Press.

[Ashby, 1952] Ashby, W. R. (1952). *Design for a Brain*. London: Chapman and Hall.

[Atkeson et al., 1997] Atkeson, C., Moore, A., and Schaal, S. (1997). Locally weighted learning. *Artificial intelligence review*, 11(1):11–73.

[Ballard, 1991] Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48(1):57–86.

[Ballard et al., 1997] Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20:723–767.

[Beer, 1995] Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1-2):173–215.

[Beer, 2003] Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243.

[Chang et al., 2003] Chang, E., Goh, K., Sychay, G., and Wu, G. (2003). Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38.

[Chemero, 2003] Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology*, 15(2).

[Collier and Ramirez-Serrano, 2009] Collier, J. and Ramirez-Serrano, A. (2009). Environment classification for indoor/outdoor robotic mapping. In *Proceedings of the IEEE Canadian Conf. on Computer and Robot Vision*, pages 276–283.

[Floreano et al., 2004] Floreano, D., Toshifumi, K., Marocco, D., and Sauser, E. (2004). Co-evolution of active vision and feature selection. *Biological Cybernetics*, 90(3):218–228.

[Fritz et al., 2006] Fritz, G., Paletta, L., Kumar, M., Dorffner, G., Breithaupt, R., and Rome, E. (2006). Visual learning of affordance based cues. In *Proceedings of the Intl. Conf. on the Simulation of Adaptive Behavior (SAB)*, volume LNCS 4095, page 52. Springer.

[Gibson, 1977] Gibson, J. (1977). The theory of affordances. *Perceiving, acting, and knowing: Toward an ecological psychology*, pages 67–82, 127.

[Gibson, 1979] Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Ass.

[Goldstein, 2002] Goldstein, E. (2002). *Sensation & Perception*.

[Goodale, 2008] Goodale, M. A. (2008). Action without perception in human vision. *Cognitive Neuropsychology*, 25(7):891–919.

[Gray and Olshen, 1997] Gray, R. and Olshen, R. (1997). Vector quantization and density estimation. In *Proceedings of the Compression and Complexity of Sequences Conference*, pages 172–193. Citeseer.

[Itti et al., 2005] Itti, L., Rees, G., and Tsotsos, J. (2005). Models of bottom-up attention and saliency. *Neurobiology of attention*, 582.

[Kim et al., 2006] Kim, D., Sun, J., Oh, S., Rehg, J., and Bobick, A. (2006). Traversability classification using unsupervised on-line visual learning for outdoor robot navigation. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006 (ICRA 2006)*, pages 518–525. IEEE.

[Kimchi, 1992] Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: A critical review. *Psychological Bulletin*, 112(1):24–38.

[Marr, 1982] Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc. New York, NY, USA.

[McVea and Pearson, 2007] McVea, D. and Pearson, K. (2007). Contextual learning and obstacle memory in the walking cat. *Integrative and Comparative Biology*.

[Mossio and Taraborelli, 2008] Mossio, M. and Taraborelli, D. (2008). Action-dependent perceptual invariants: From ecological to sensorimotor approaches. *Consciousness and Cognition*, 17(4):1324–1340.

[Nolfi, 2005] Nolfi, S. (2005). *Handbook of Categorization in Cognitive Science*, chapter Categories Formation in Self-Organizing Embodied Agents, pages 869–889. Elsevier.

[Oliva and Schyns, 2000] Oliva, A. and Schyns, P. (2000). Diagnostic Colors Mediate Scene Recognition* 1. *Cognitive Psychology*, 41(2):176–210.

[Oliva and Torralba, 2007] Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527.

[Pearson, 2000] Pearson, K. (2000). Neural adaptation in the generation of rhythmic behavior. *Annual review of physiology*, 62(1):723–753.

[Pfeifer and Scheier, 1999] Pfeifer, R. and Scheier, C. (1999). *Understanding intelligence*. MIT Press, Massachussetts.

[Reynolds and Bronstein, 2004] Reynolds, R. and Bronstein, A. (2004). The moving platform aftereffect: limited generalization of a locomotor adaptation. *Journal of neurophysiology*, 91(1):92.

[Santana et al., 2010] Santana, P., Santos, C., Chaínho, D., Correia, L., and Barata, J. (2010). Predicting Affordances from Gist. In *Proceedings of the International Conference on Simulation of Adaptive Behavior (SAB 2010)*, pages 325–334, Paris. Springer.

[Scheier et al., 1998] Scheier, C., Pfeifer, R., and Kuniyoshi, Y. (1998). Embedded neural networks: exploiting constraints. *Neural Networks*, (11):1551–1596.

[Schyns and Oliva, 1994] Schyns, P. and Oliva, A. (1994). From Blobs to Boundary Edges. *Psychological Science*, 5(4):195–200.

[Siagian and Itti, 2005] Siagian, C. and Itti, L. (2005). Gist: A Mobile Robotics Application of Context-Based Vision in Outdoor Environment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops.*, page 88. IEEE.

[Siagian and Itti, 2007] Siagian, C. and Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on PAMI*, pages 300–312.

[Simoncelli and Freeman, 1995] Simoncelli, E. and Freeman, W. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings of the 1995 International Conference on Image Processing*, volume 3, pages 444–447.

[Slocum et al., 2000] Slocum, A., Downey, D., and Beer, R. (2000). Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In *Proceedings of the Intl. Conf. on Simulation of Adaptive Behavior (SAB)*, volume 6, pages 430–439.

[Sporns and Lungarella, 2006] Sporns, O. and Lungarella, M. (2006). Evolving coordinated behavior by maximizing information structure. In *Proceedings of the Intl. Conf. on the Simulation and Synthesis of Living Systems (ALife X)*, pages 3–7.

[Thelen and Smith, 1996] Thelen, E. and Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action*. The MIT Press.

[Thorpe et al., 1996] Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

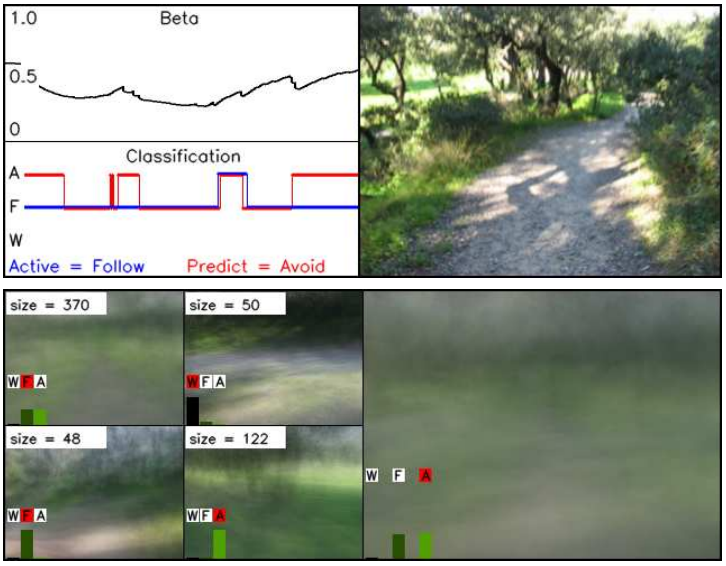
[Torralba et al., 2003] Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *Proceedings of the IEEE ICCV*, pages 273–280.

[Vailaya et al., 1999] Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H. J. (1999). Content-based hierarchical classification of vacation images. In *IEEE International Conference on Multimedia Computing and Systems, 1999.*, volume 1, pages 518–523 vol.1.

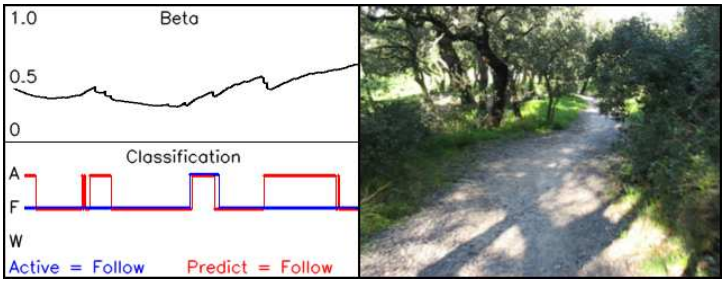
Appendix A

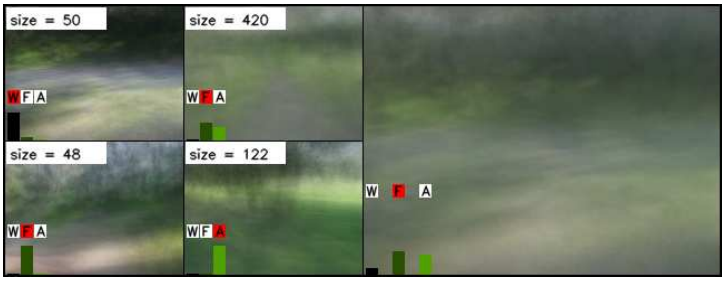
Additional Results

Frame #3135

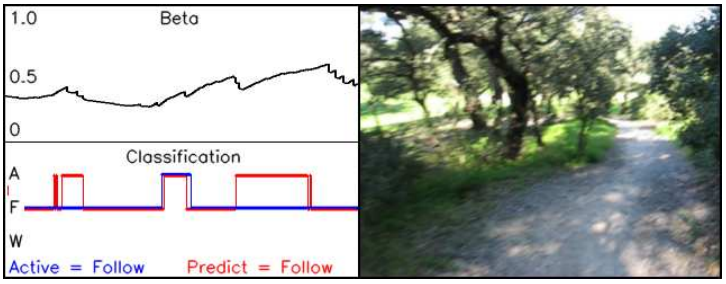


Frame #3160





Frame #3185



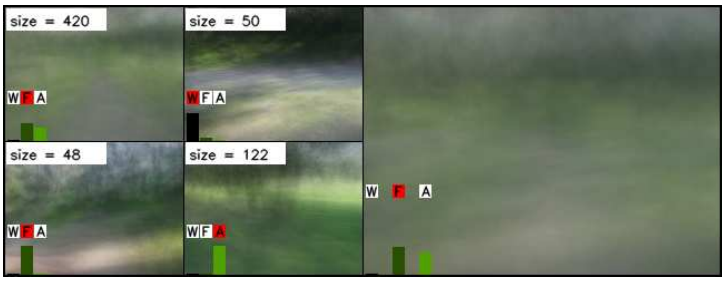
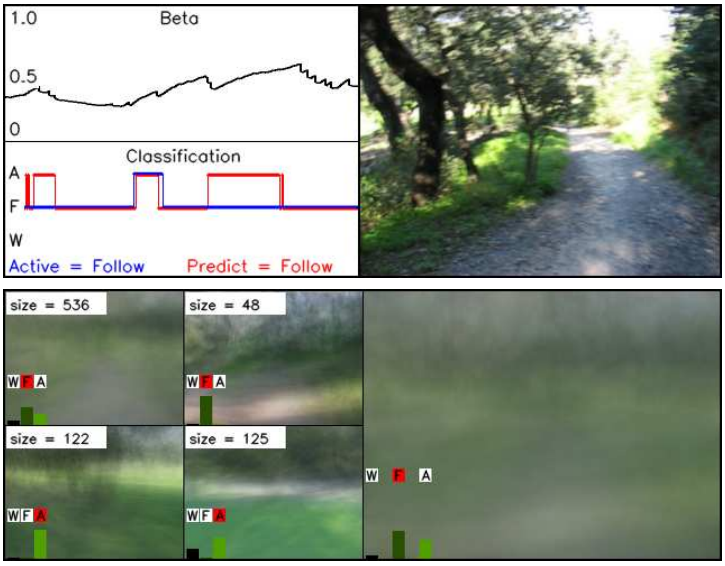
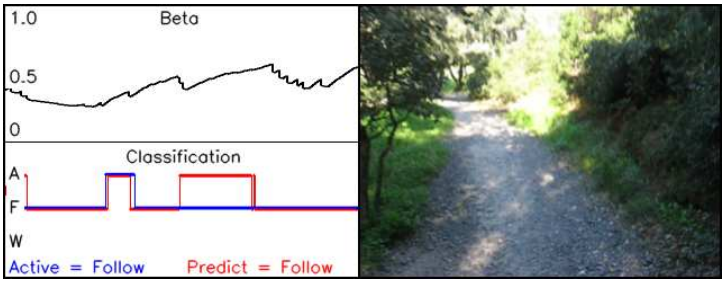


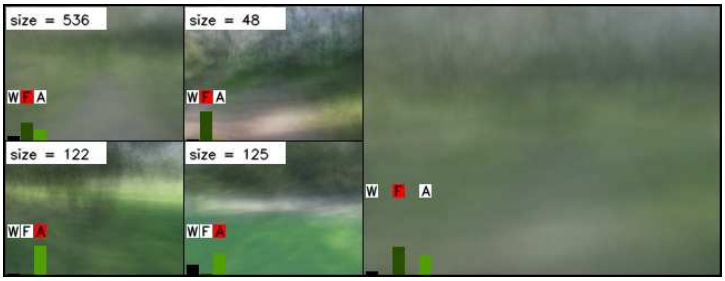
Figure A.1: Frames from location a in Fig.4.1.

Frame #3210

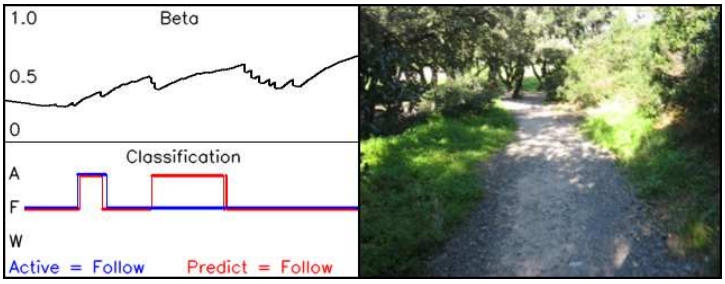


Frame #3235





Frame #3260



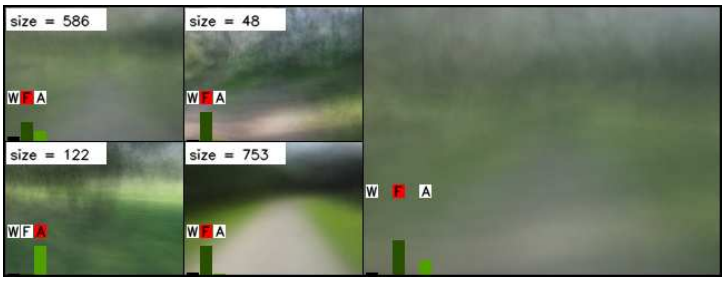
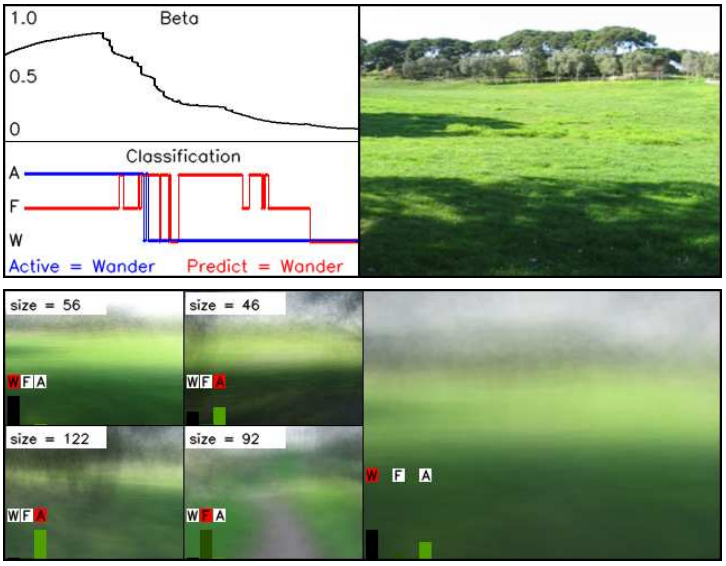
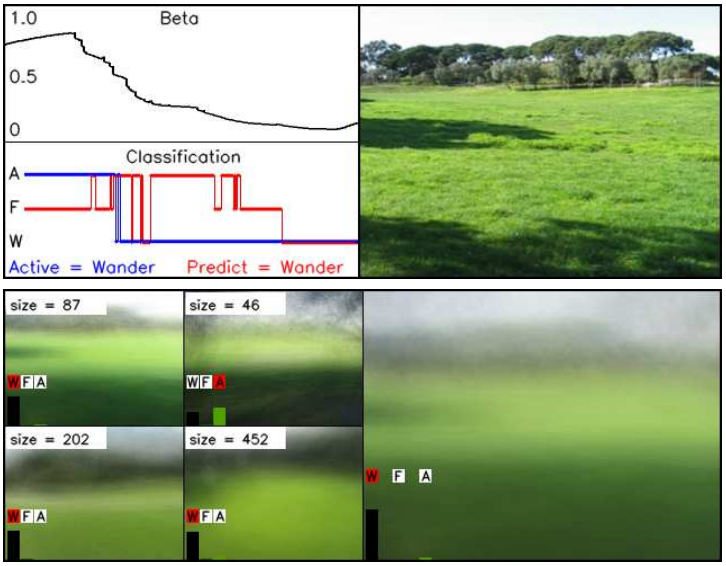


Figure A.2: Frames from location a in Fig.4.1.

Frame #3939



Frame #3964



Frame #3989

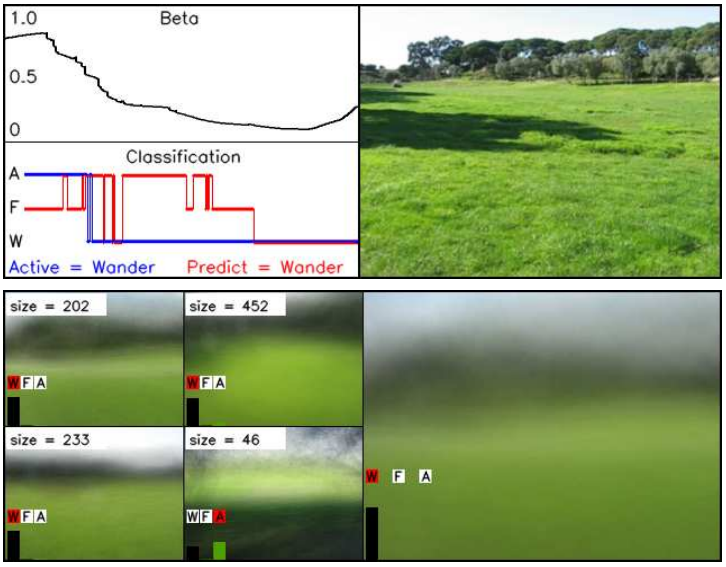
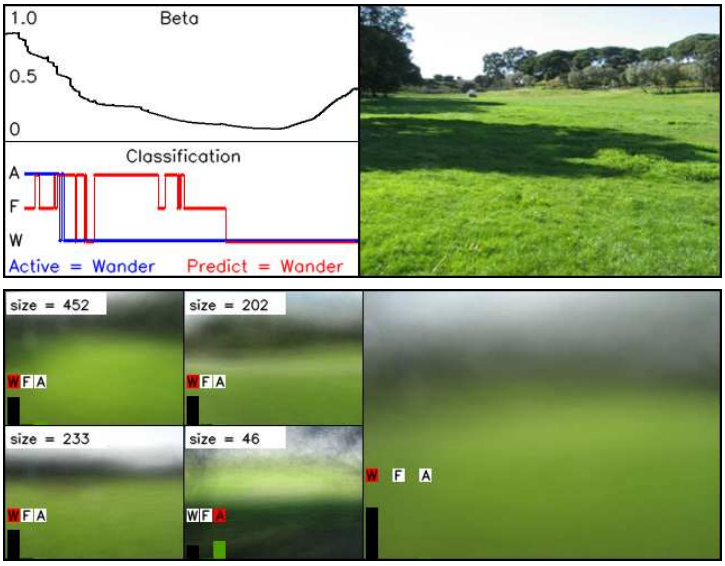
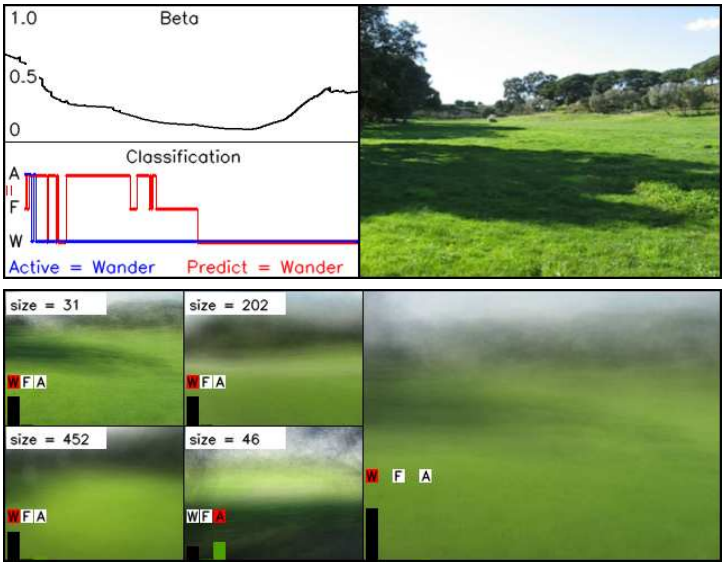


Figure A.3: Frames from location b in Fig.4.1.

Frame #4014



Frame #4039



Frame #4064

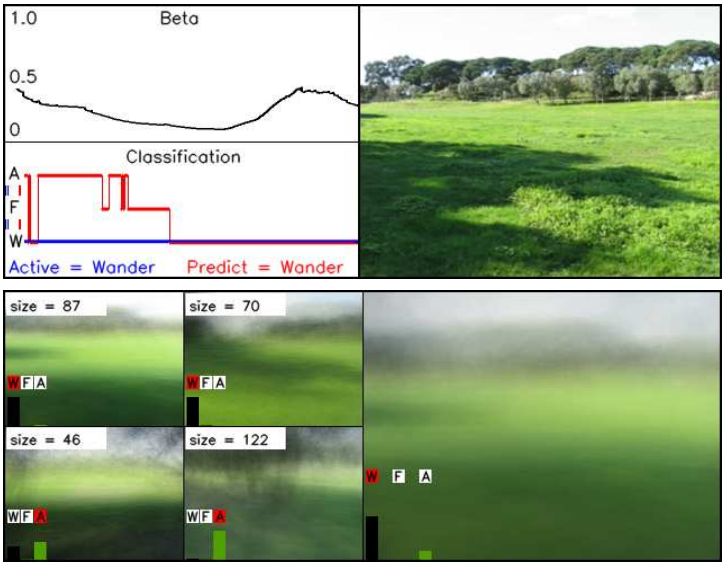
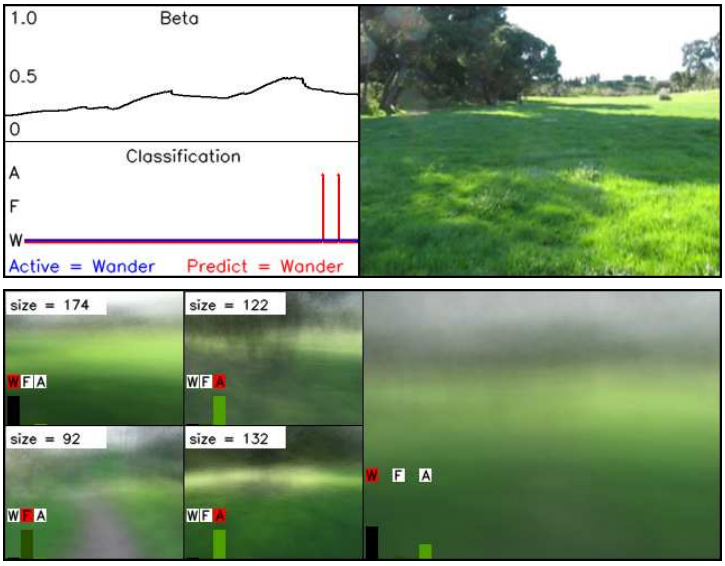
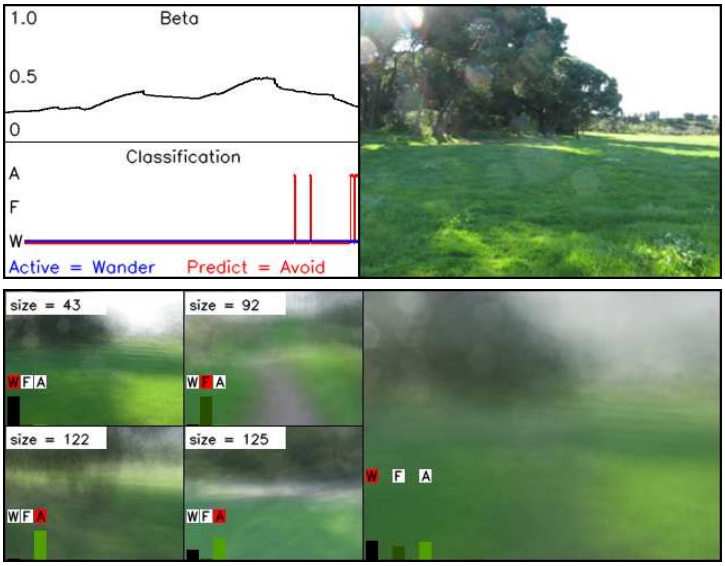


Figure A.4: Frames from location b in Fig.4.1.

Frame #4591



Frame #4616



Frame #4641

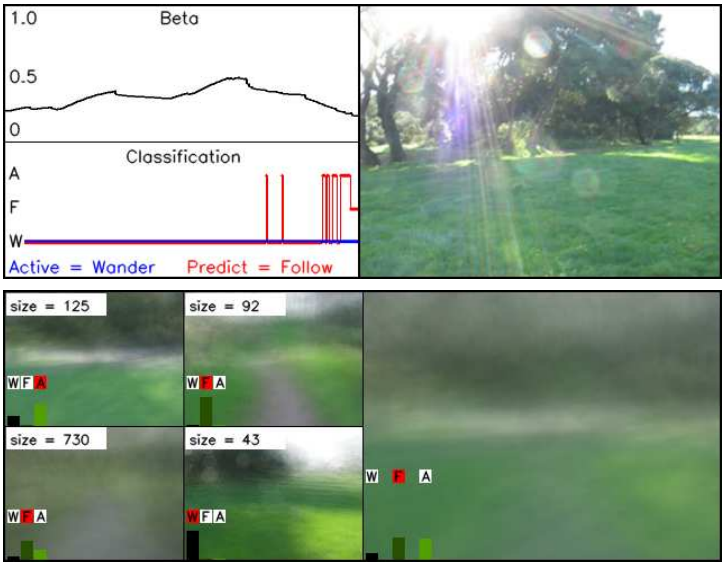
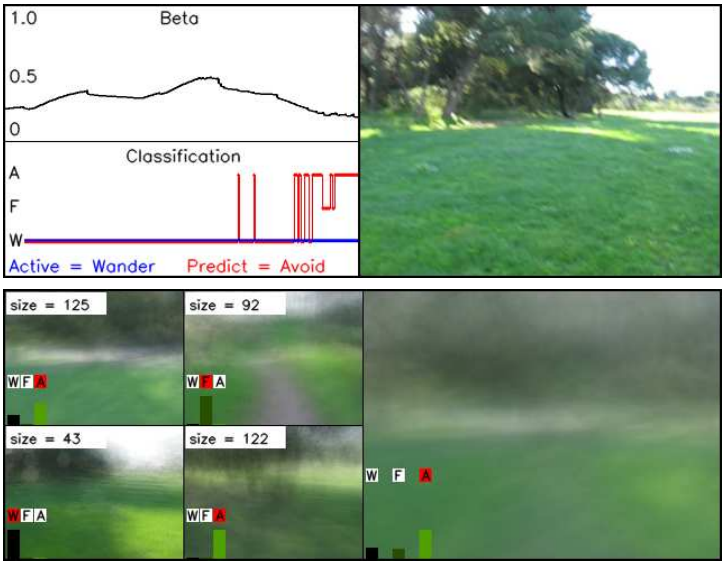
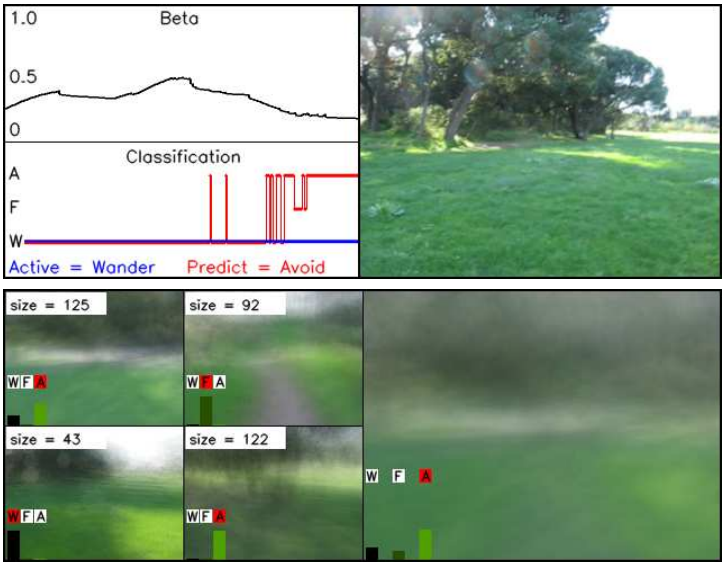


Figure A.5: Frames from location d in Fig.4.1.

Frame #4666



Frame #4691



Frame #4716

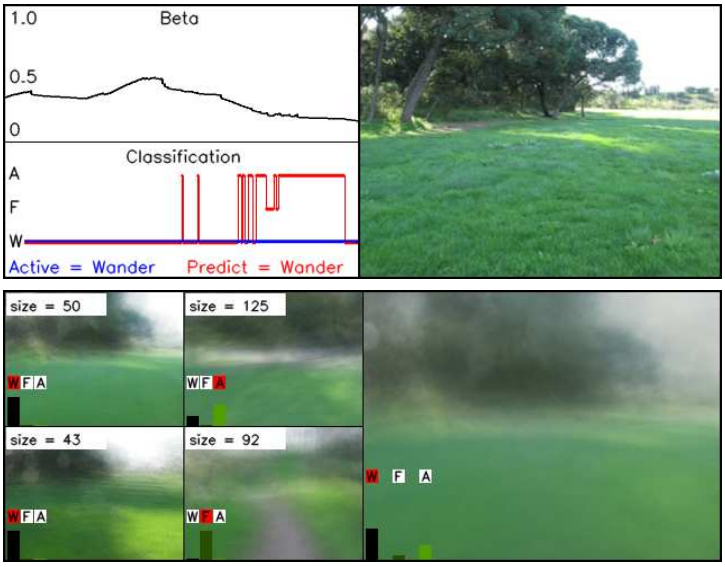
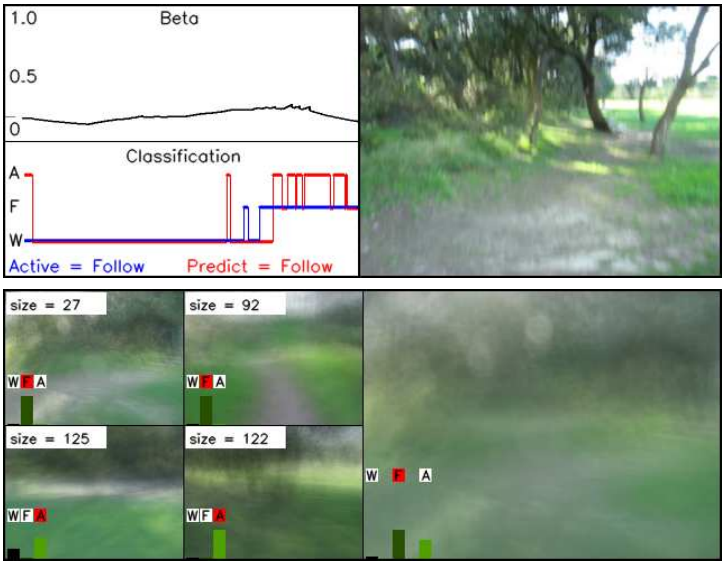
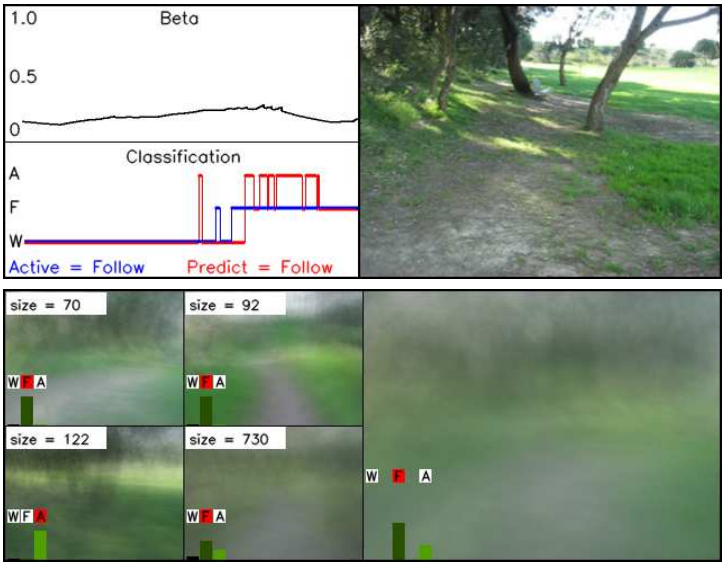


Figure A.6: Frames from location d in Fig.4.1.

Frame #4994



Frame #5019



Frame #5044

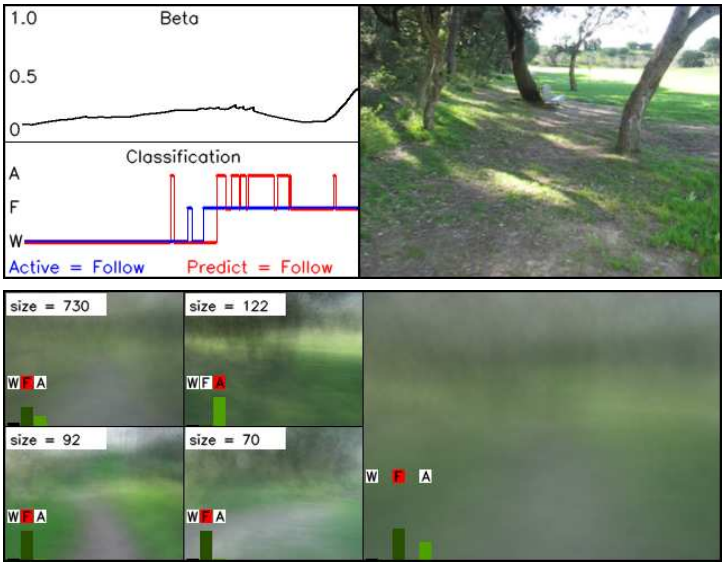
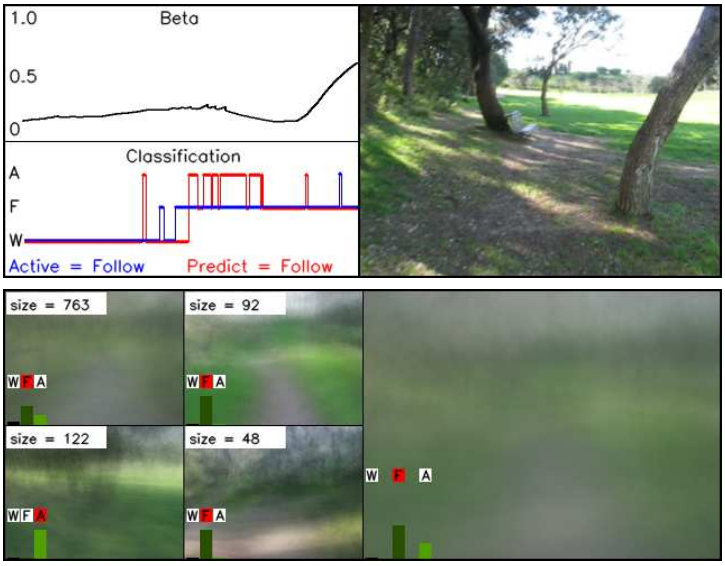


Figure A.7: Frames from location e in Fig.4.1.

Frame #5069



Frame #5094

Frame #5119

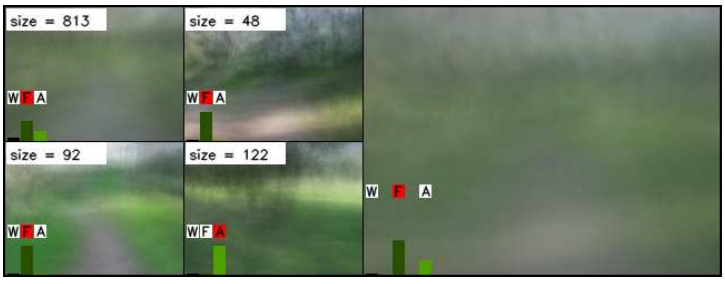
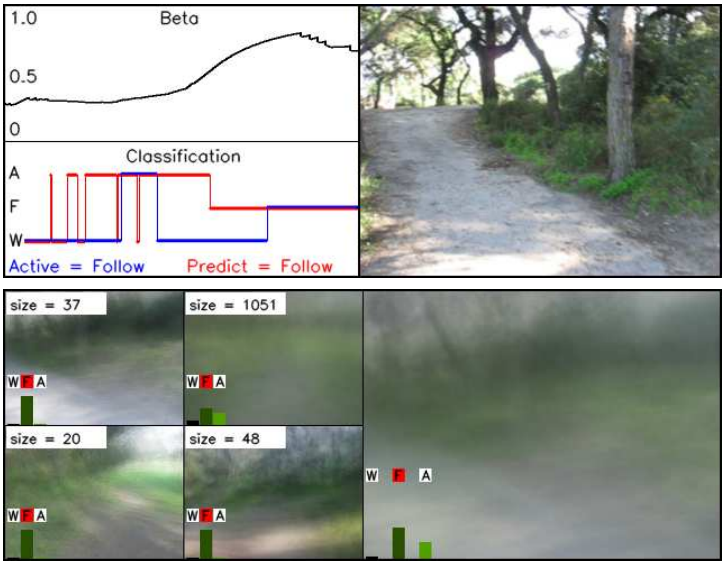
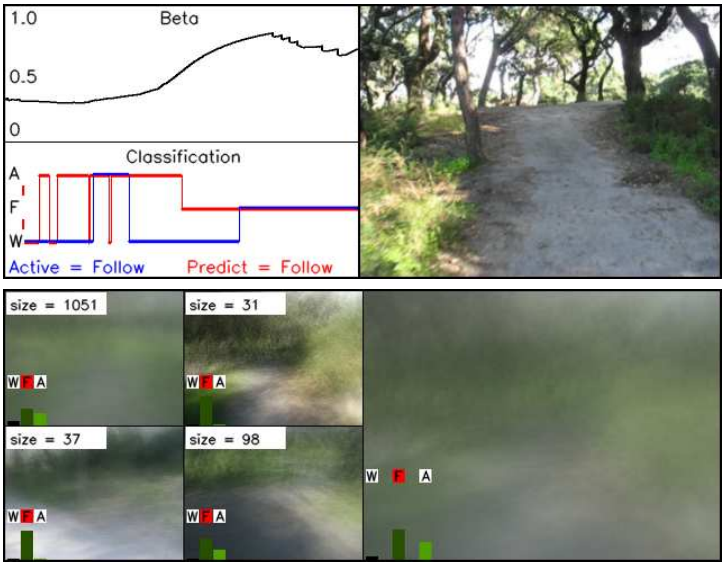


Figure A.8: Frames from location e in Fig.4.1.

Frame #5757



Frame #5782



Frame #5807

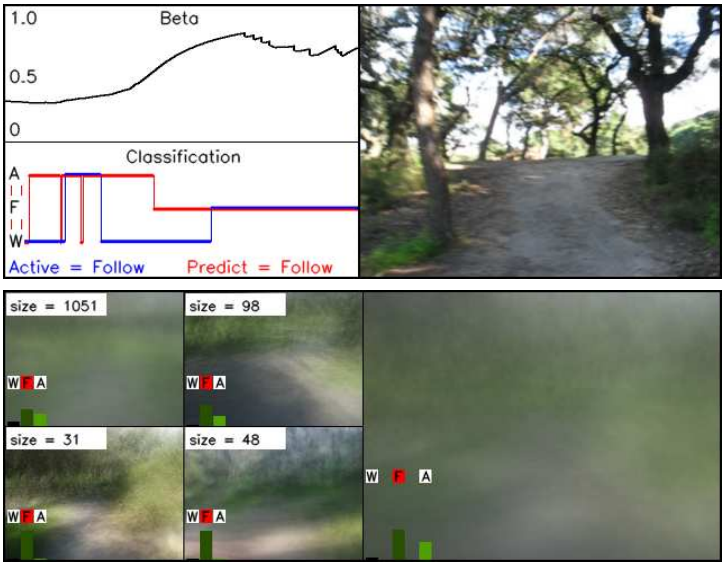
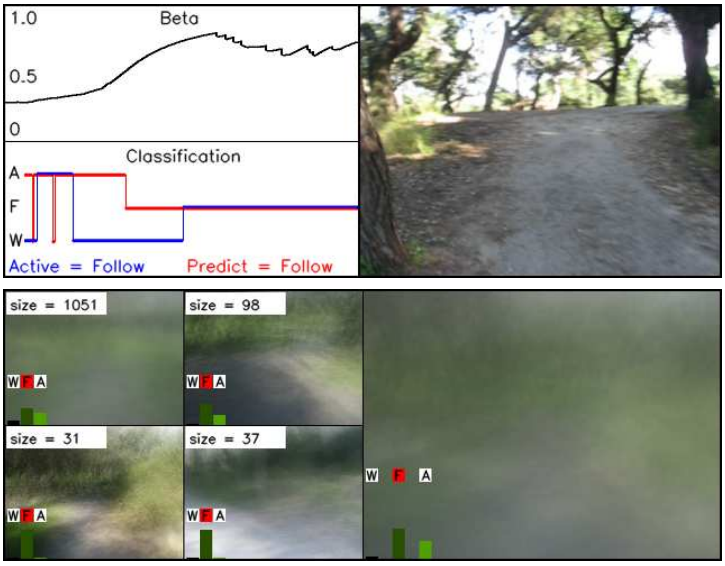
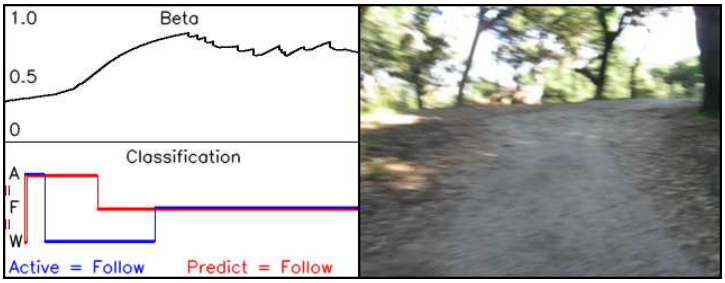


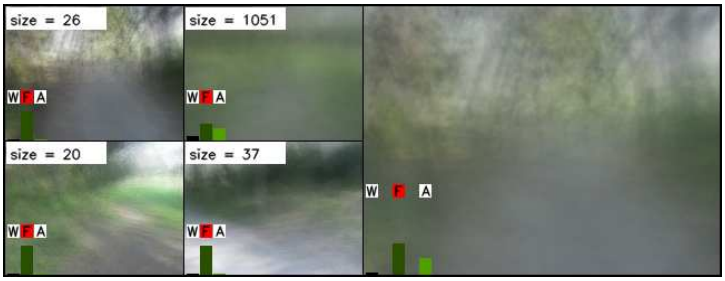
Figure A.9: Frames from location `i` in Fig.4.1.

Frame #5832



Frame #5857





Frame #5882

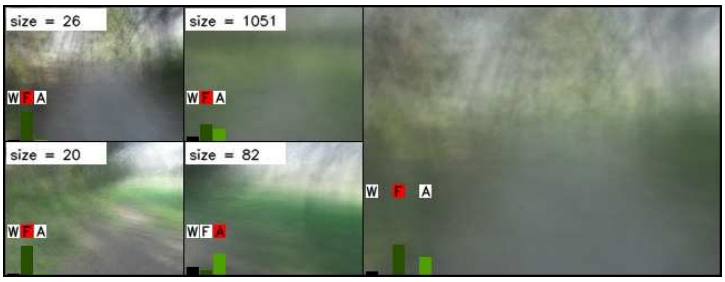
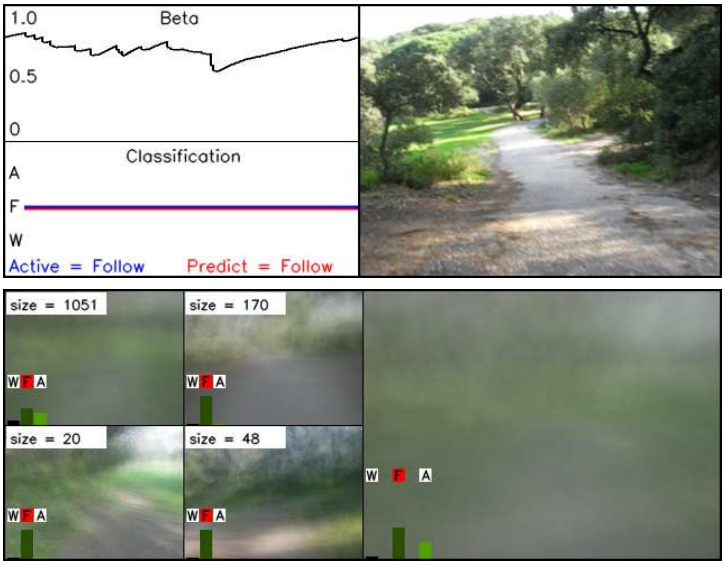
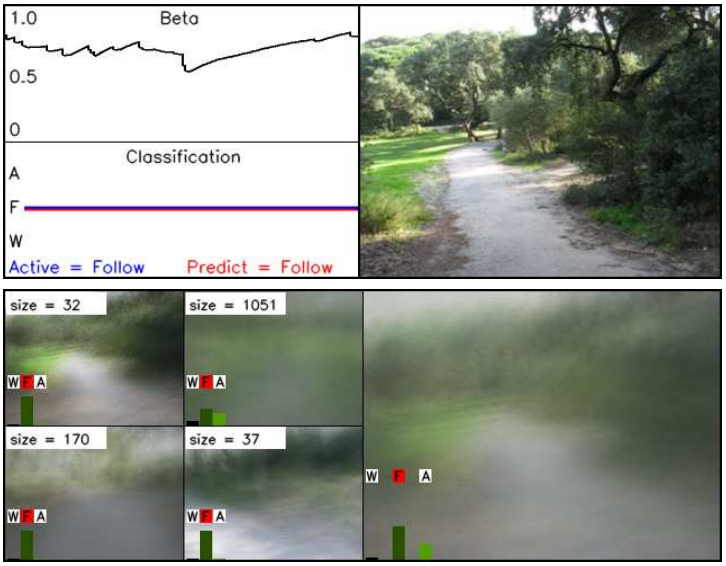


Figure A.10: Frames from location `i` in Fig.4.1.

Frame #6002



Frame #6027



Frame #6052

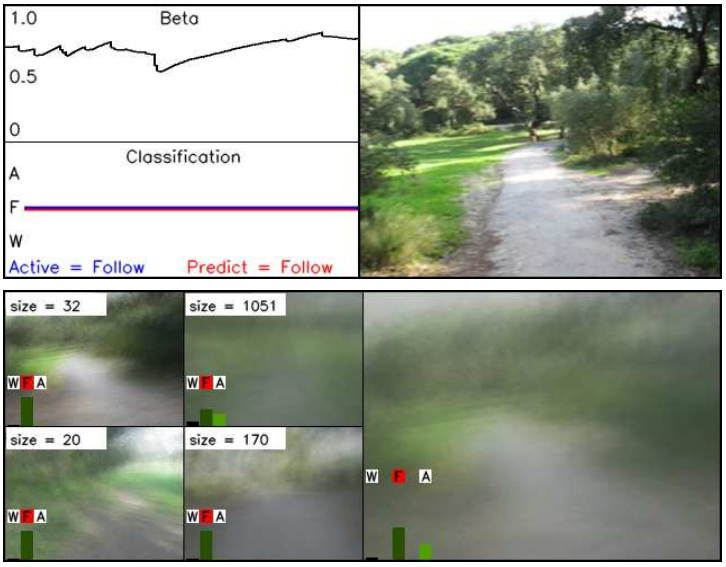
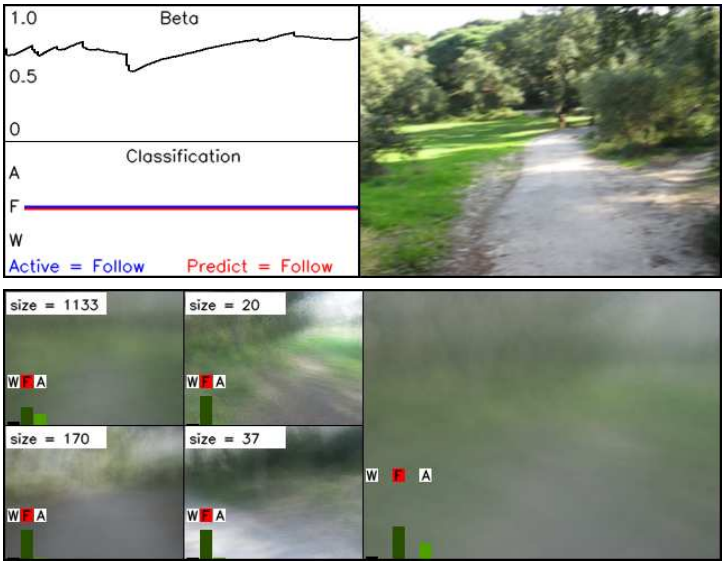
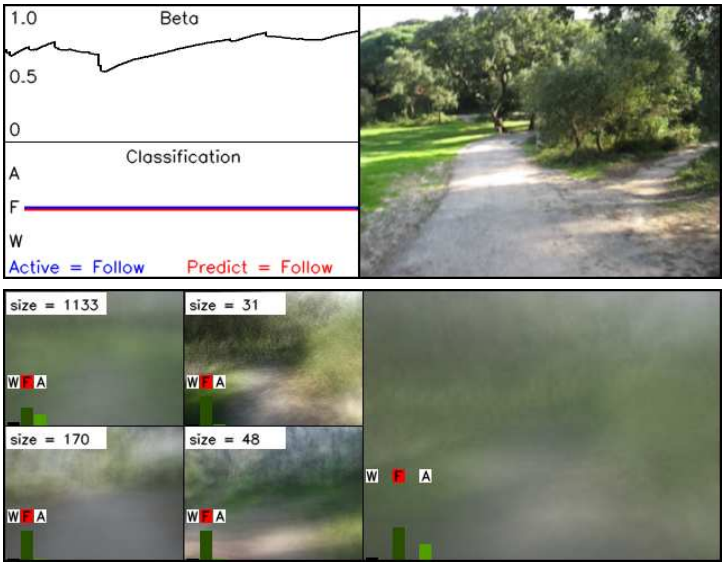


Figure A.11: Frames from location `j` in Fig.4.1.

Frame #6077



Frame #6102



Frame #6127

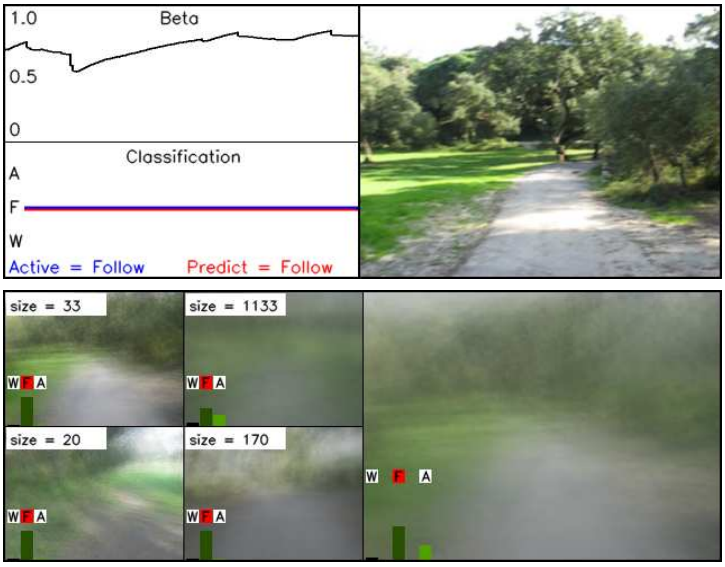
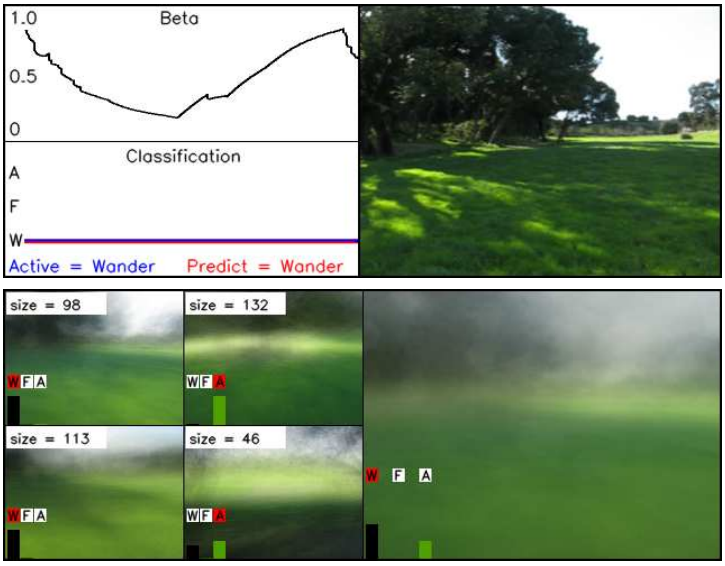
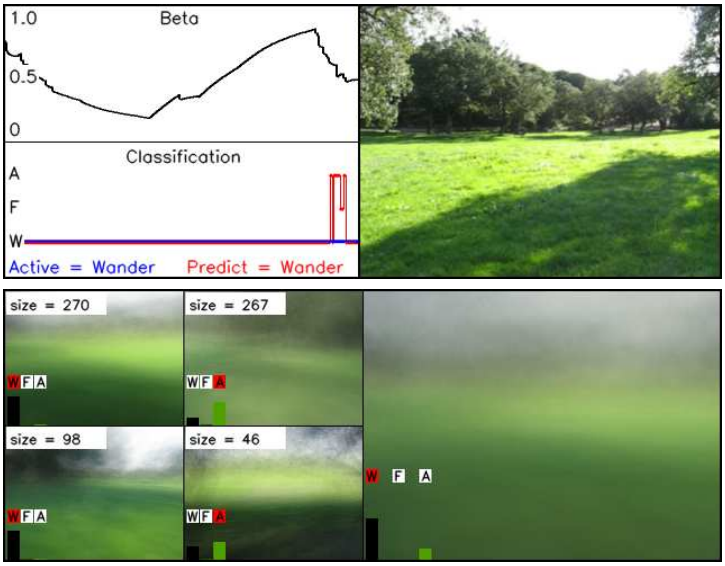


Figure A.12: Frames from location j in Fig.4.1.

Frame #7595



Frame #7620



Frame #7645

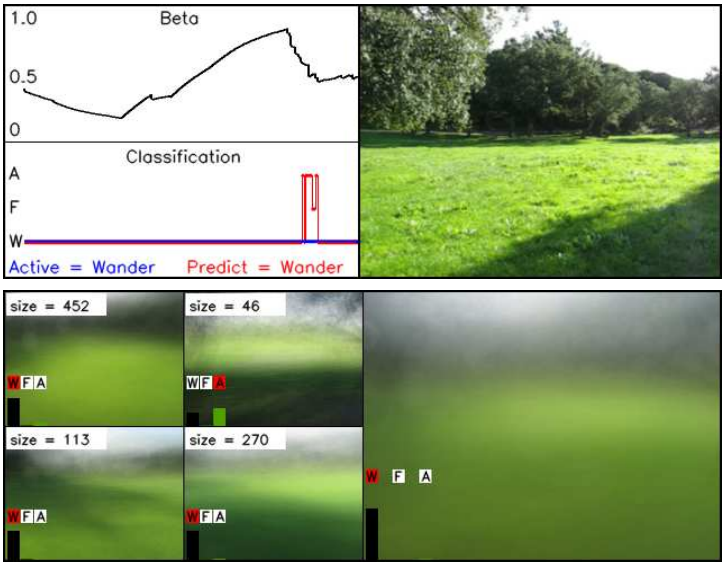
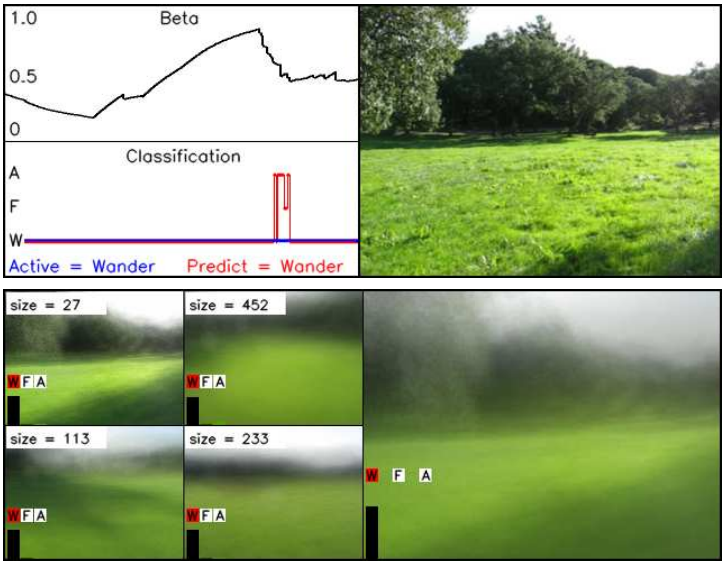
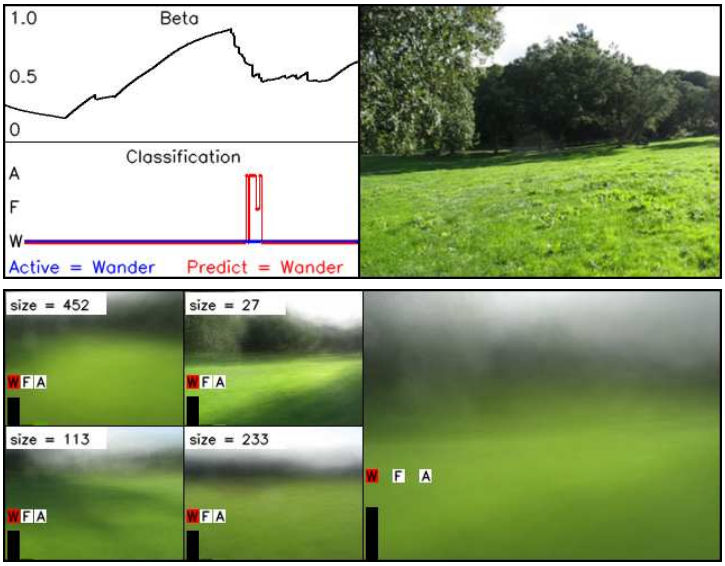


Figure A.13: Frames from location k in Fig.4.1.

Frame #7670



Frame #7695



Frame #7720

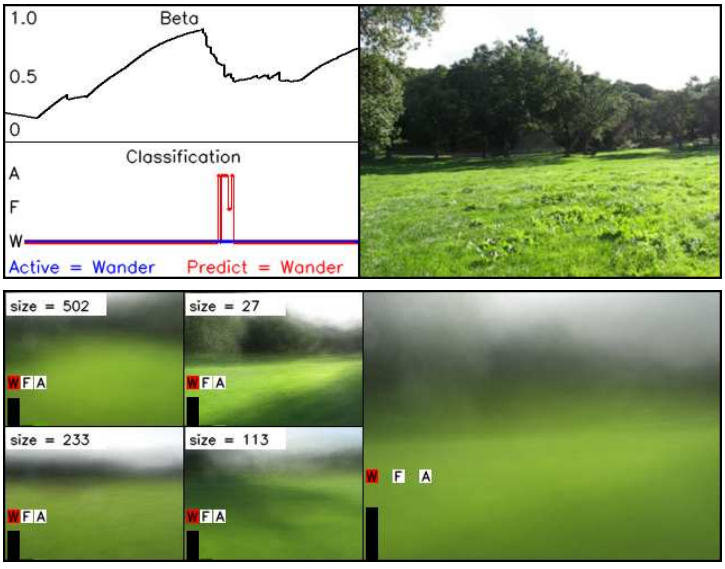
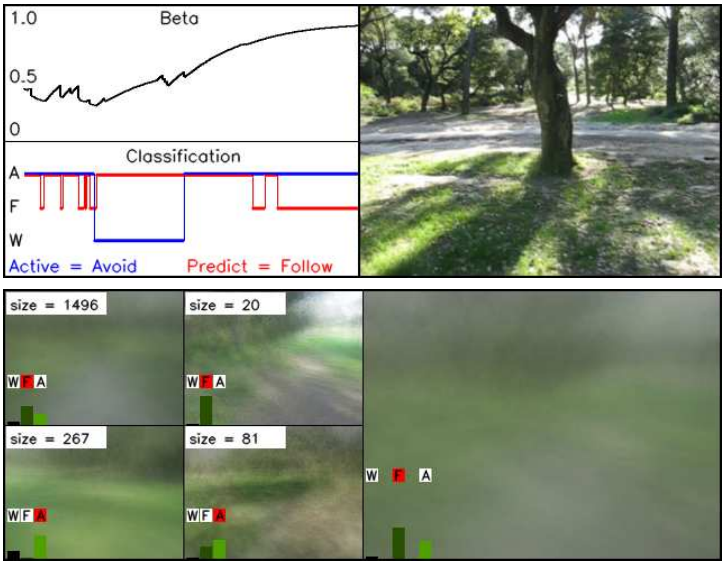
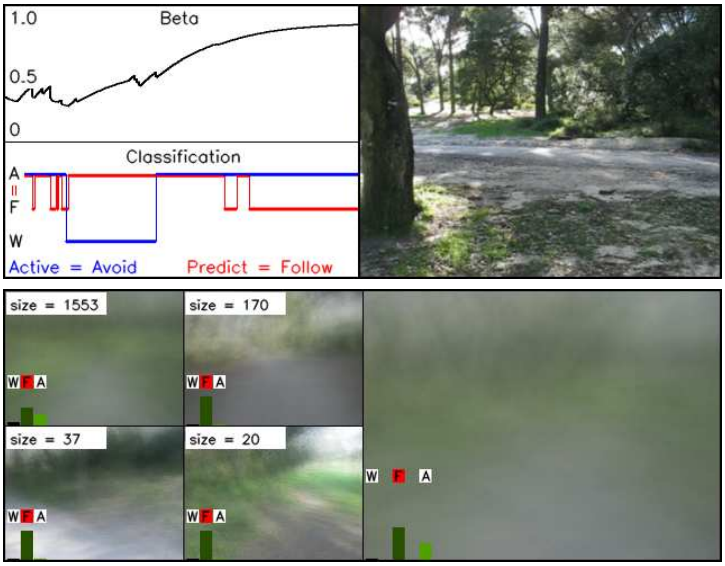


Figure A.14: Frames from location k in Fig.4.1.

Frame #8707



Frame #8732



Frame #8757

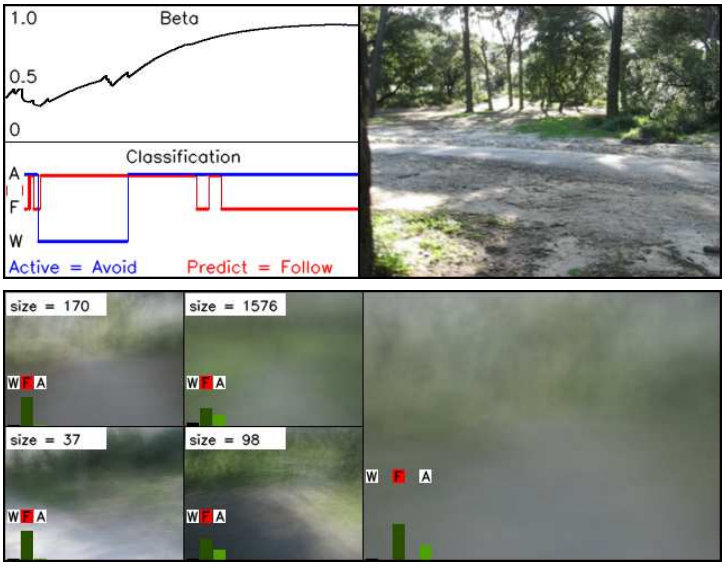
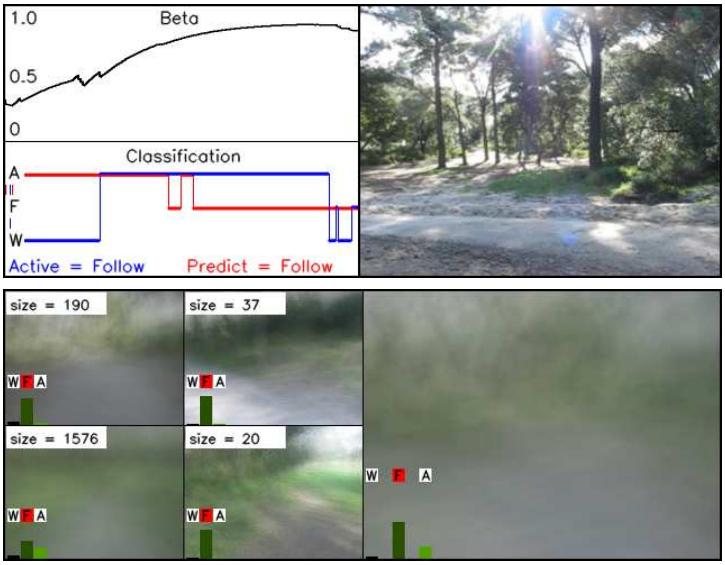
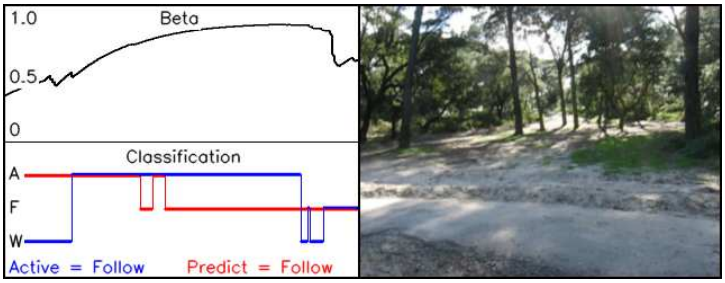


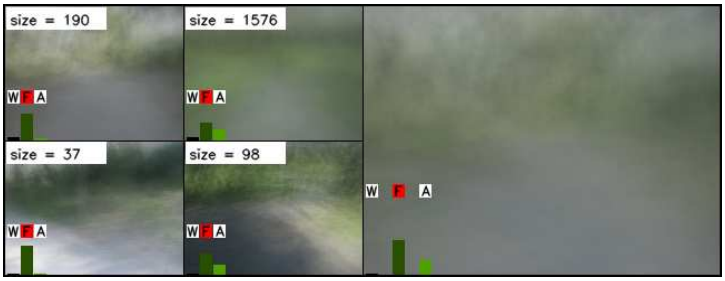
Figure A.15: Frames from location l in Fig.4.1.

Frame #8782

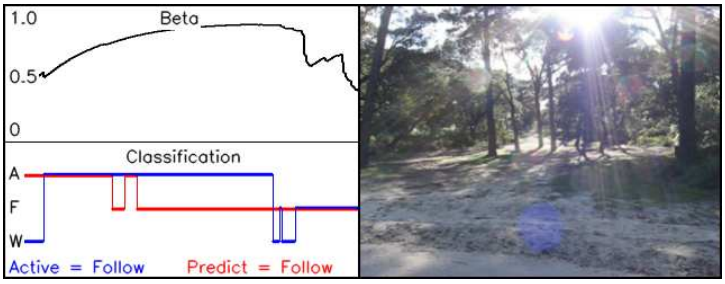


Frame #8807





Frame #8832



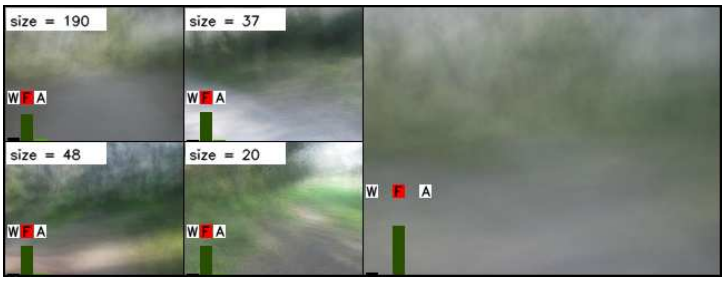


Figure A.16: Frames from location l in Fig.4.1.